MỤC LỤC

DANH MỤC HÌNH ẢNH, BẢNG BIỂU VÀ BIỂU ĐỒ

1. Danh mục hình ảnh

2. Danh mục bảng biểu

3. Danh mục biểu đồ

CHƯƠNG 1: TỔNG QUAN

1. Lý do chọn đề tài

Cùng với những tính cách mạnh mẽ ở các thành viên trong nhóm, kết hợp với sự hứng thú trải dài ở vô vàn những chủ đề khác nhau, để tìm được điểm giao thoa không những phải phù hợp với mối quan tâm của nhóm mà còn cần đáp ứng được yêu cầu của bài đồ án là một câu chuyện hết sức nan giải. Thế nhưng sau một khoảng thời gian dài cùng với những trận đấu trí khốc liệt không hồi kết, nhóm chúng em cuối cùng đã tìm được cho mình một chủ đề phù hợp với xu thế hiện tại: Thương mại điện tử. E-commerce là đề tài rất phù hợp cho bài đồ án môn Khoa Học Dữ Liệu vì nó có nhiều lý do hấp dẫn để nghiên cứu và phân tích. Đầu tiên, đây là một lĩnh vực có tính ứng dụng cao trong thực tế bởi khi phân tích dữ liệu E-commerce có thể giúp các doanh nghiệp hiểu rõ hơn về khách hàng, sản phẩm, xu hướng mua sắm và từ đó đưa ra các *chiến lược kinh doanh* hiệu quả. Thứ hai, E-commerce là một lĩnh vực có rất nhiều dữ liệu khác nhau, từ thông tin sản phẩm, thông tin khách hàng, đơn hàng, thanh toán và giao nhận và việc này đòi hỏi phải xử lý khối dữ liệu khác nhau này sẽ giúp cho việc nghiên cứu trở nên thú vị và phong phú hơn. Vì vậy, việc lựa chọn đề tài E-commerce cho bài đồ án môn Khoa học dữ liệu là một sự lựa chọn đầy tiềm năng. Nhóm đặt sự ưu tiên lên hàng đầu trong việc nghiên cứu và trau dồi khả năng sử dụng các công cụ cần thiết mà đã được thầy hướng dẫn học tập trong suốt quá trình vừa qua. Bên cạnh đó, chủ đề còn đáp ứng được sự quan tâm của nhóm đối với ngành học hiện tại là Hệ thống thông tin kinh doanh và định hướng công việc Business Analyst cho sau này. Dù đồ án chỉ nằm trong một phạm vi nhỏ trong vị trí công việc, nhưng bằng cách áp dụng các công cụ như Excel, Orange,... có thể giúp cho Business Analyst có cái nhìn sâu hơn thông qua việc chia tách vấn đề thành các cụm nhỏ lẻ và phân tích insights từ đó - công việc vô cùng quan trọng trước khi giúp doanh nghiệp đưa ra giải pháp.

2. Mục tiêu nghiên cứu

Nghiên cứu sẽ cung cấp thông tin của đối tượng cần tìm hiểu, sử dụng các công cụ phân tích để đưa ra kết luận cụ thể, cũng như đưa ra hướng đi hay giải pháp cho bất kỳ doanh nghiệp hay tổ chức để tìm kiếm và dự đoán khả năng rời khỏi hệ thống của khách hàng mới đồng thời giữ chân các đối tượng khách hàng cũ

Có 3 mục tiêu chính của đề tài cũng ứng với 3 bài toán cần giải quyết của bài nghiên cứu.

Bài toán phát hiện điểm đặc thù của dữ liệu: Sử dụng các công cụ thống kê thông dụng như Pivot Table, các hàm của Excel, Orange và các dạng lược đồ, biểu đồ để phát hiện, thể hiện các điểm đặc thù của dữ liệu và mối quan hệ giữa chúng.

Ứng dụng bài toán phân lớp để dự đoán khả năng rời đi của khách hàng và phân tích lý do, tìm hiểu vấn đề đằng sau là gì.

Dùng phương pháp Hierarchical Clustering để thực hiện và phân loại khách hàng làm các cụm để nhận thấy rõ đặc điểm của khách hàng.

3. Đối tượng và phạm vi nghiên cứu

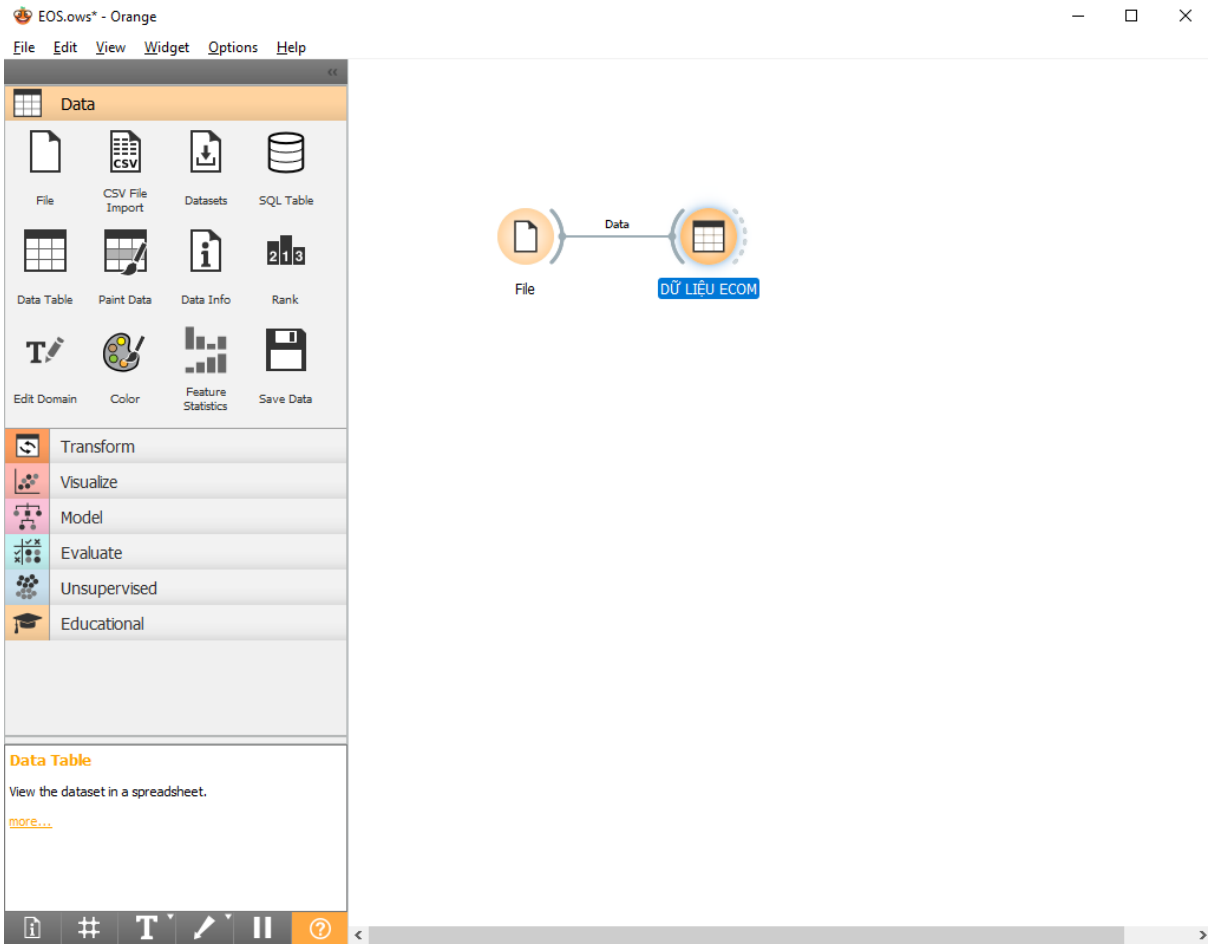
Đồ án này sẽ tập trung nghiên cứu hành vi và thông tin của tệp khách hàng của một sàn thương mại điện tử được lấy thông tin, dữ liệu cũng như số liệu tại Kaggle.

CHƯƠNG 2: PHÂN TÍCH DỮ LIỆU E-COMMERCE CHURN RATE

Nội dung chương: Ứng dụng vào bài toán thực tế những kiến thức đã học và kiến thức liên quan đến phần mềm Orange để bước đầu phân tích dữ liệu, lựa chọn các phương pháp phù hợp để tiến hành Tiền xử lý dữ liệu (xử lý dữ liệu bị thiếu/lỗi; phân tách dữ liệu; xác định các loại biến), Mô tả và Thống kê mô tả dữ liệu.

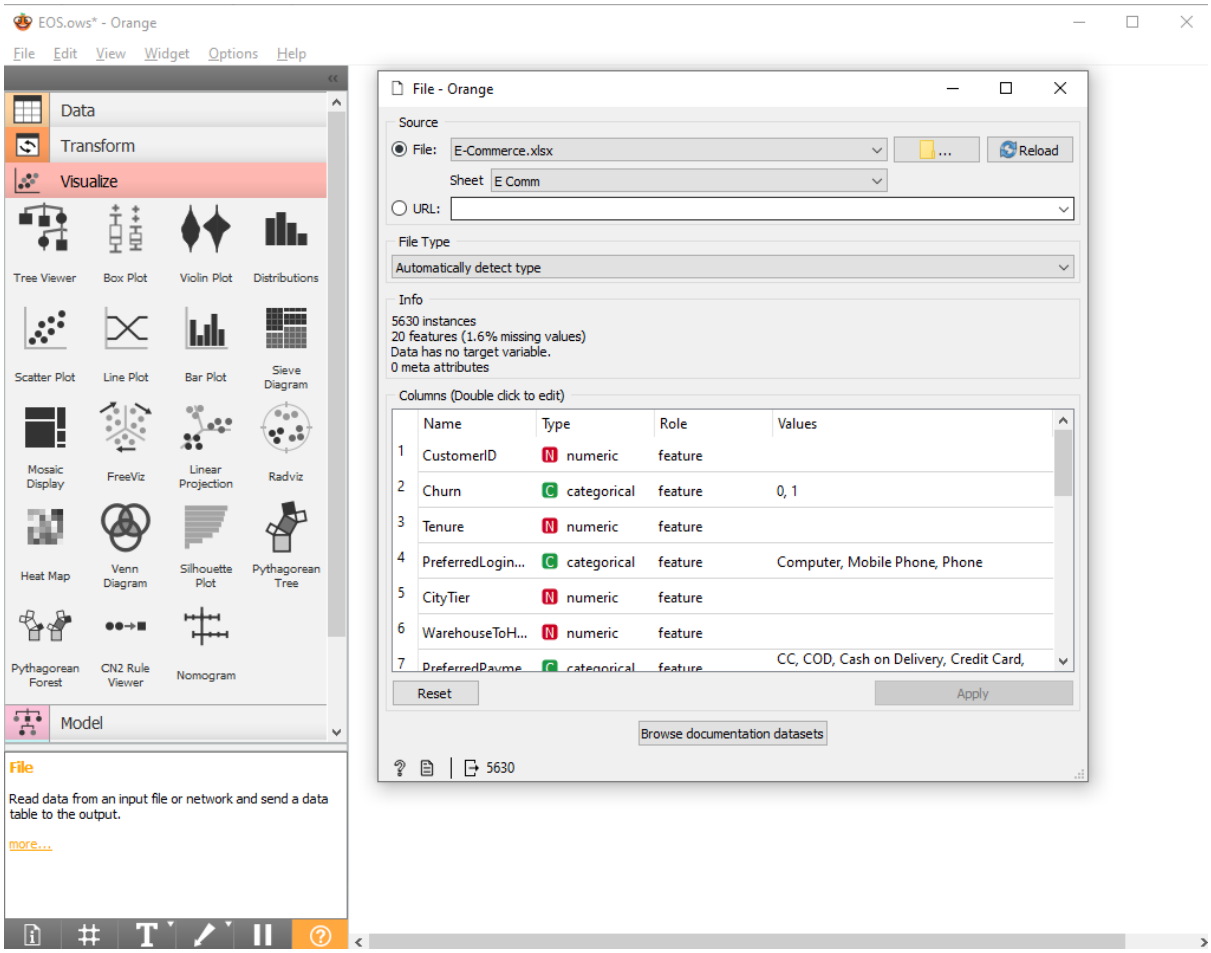
1. Tiền xử lý Dữ liệu E-Commerce Churn Rate

- Dữ liệu E-commerce Churn Rate được nhóm tìm kiếm và chọn lọc từ trang web https://www.kaggle.com/datasets. Sau đó, tiến hành xử lý dữ liệu gồm các bước tại phần mềm Orange:



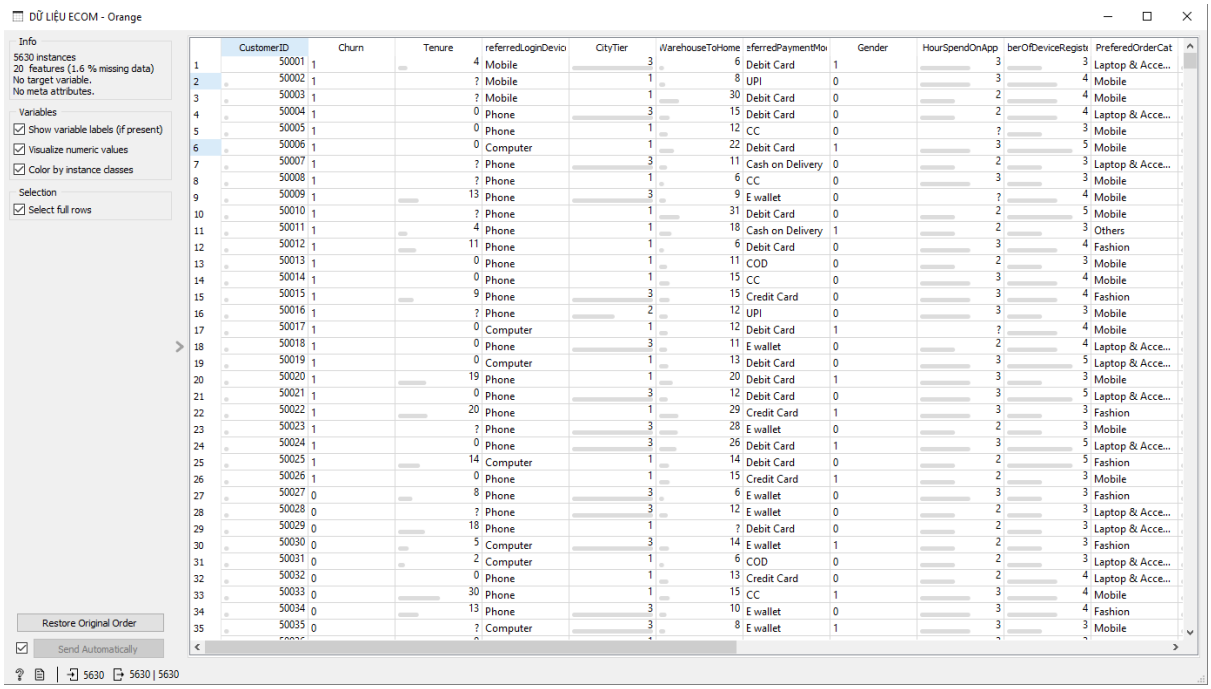
Hình 1. Các bước tiến hành xử lý dữ liệu

Bước 1: Nạp dữ liệu E-Commerce: mở file chọn E-Commerce.



Hình 2. Nạp dữ liệu E-Commerce vào trong File

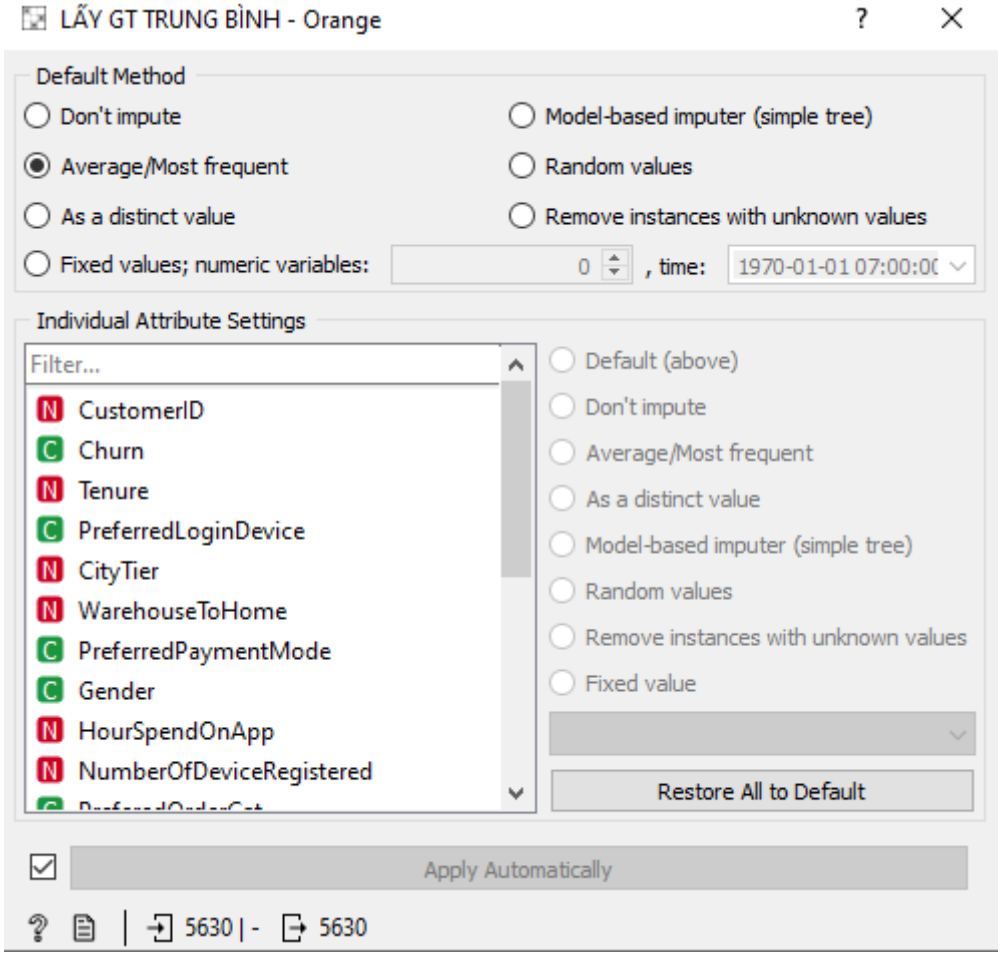
Bước 2: Quan sát dữ liệu: mở Data Table và nối File vào Data Table. Quan sát dữ liệu, ta thấy có 1.6% dữ liệu bị thiếu. Do đó, ta tiến hành xử lý dữ liệu bị thiếu đó.



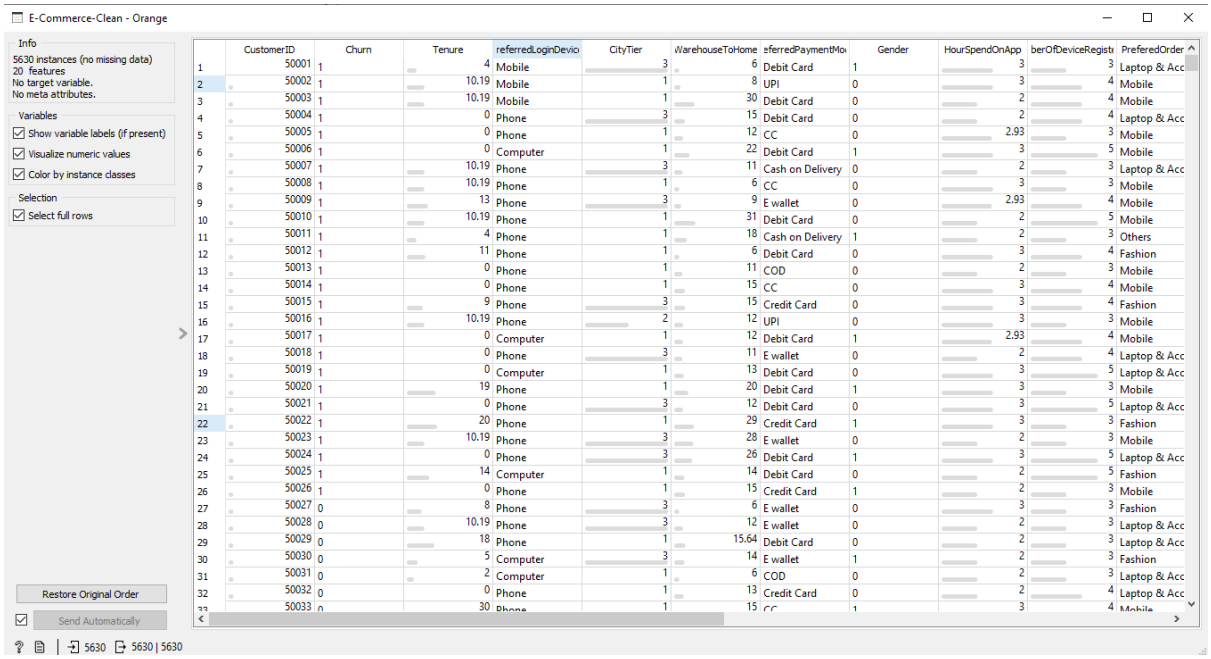
Hình 3. Kết quả kiểm tra dữ liệu bị thiếu

- Sử dụng công cụ Feature Statistics thấy được các thuộc tính sau đây bị mất dữ liệu:
    + Tenure: 264 ô dữ liệu (5%).
    + WareHouseToHouse: 251 ô dữ liệu (4%).
    + HourSpendOnApp: 255 ô dữ liệu (5%).
    + OrderAmountHikeFromLastYear: 265 ô dữ liệu (5%).
    + CouponUsed: 256 ô dữ liệu (5%).
    + OrderCount: 258 ô dữ liệu (5%).
    + DaySinceLastOrder: 307 ô dữ liệu (5%).

Bước 3: Xử lý các dữ liệu bị thiếu: nhóm sử dụng công cụ Preprocess để thay thế các dữ liệu bị thiếu này bằng giá trị trung bình của các giá trị trong thuộc tính đó.



Hình 4. Chọn điều kiện thay thế dữ liệu bị thiếu bằng giá trị trung bình



Hình 5. Bảng dữ liệu mới sau khi Preprocess Impute

Bước 4: Sau khi hoàn thành tiền xử lý bị thiếu, nhóm tiến hành lưu dữ liệu mới thành file "E-Commerce-Clean.xlsx".

Bước 5: Phân tách dữ liệu:

- Lọc từ dữ liệu "E-Commerce-Clean.xlsx", nhóm đã sử dụng công cụ Data Sampler tách dữ liệu khảo sát ban đầu thành hai file riêng biệt để thực hiện việc phân lớp dữ liệu như sau: Sử dụng 70% dữ liệu ban đầu để làm dữ liệu mẫu huấn luyện mô hình phân lớp dữ liệu (E-Commerce-Training.xlsx). Và sử dụng 30% dữ liệu còn lại để làm dữ liệu dự báo cho nghiên cứu (E-Commerce-Forecast.xlsx).
- Xác định biến độc lập và biến phụ thuộc:
  + Biến phụ thuộc là "Churn".
  + Biến độc lập là các thuộc tính còn lại.
  + Biến định danh "CustomerID", "Selected" là Skip.

2. Mô tả dữ liệu

STT	Thuộc tính	Ý nghĩa	Kiểu dữ liệu	Role
1	**CustomerID**	Mã khách hàng.	Số Thực	Skip
2	**Churn**	Khách hàng rời bỏ dịch vụ.	Số Nguyên (Biến Định Danh). 0 - Không rời bỏ 1 - Rời bỏ	Target

3	**Tenure**	Thời gian khách hàng gắn bó với tổ chức (tháng)	Số Thực	Feature
4	**PreferredLoginDevice**	Thiết bị đăng nhập ưa thích của khách hàng.	Chuỗi	Feature
5	**CityTier**	Cấp Thành Phố.	Số Nguyên	Feature
6	**WarehouseToHome**	Khoảng cách giữa nhà kho đến nhà khách hàng (km).	Số Thực	Feature
7	**PreferredPaymentMode**	Hình thức thanh toán ưa thích của khách hàng.	Chuỗi	Feature
8	**Gender**	Giới tính khách hàng.	Số Nguyên (Biến Định Danh). 0 - Nam. 1 - Nữ.	Feature
9	**HourSpendOnApp**	Thời gian khách hàng dành ra để lướt app hoặc web.	Số Thực	Feature
10	**NumberOfDeviceRegistered**	Tổng số thiết bị mà một khách hàng đăng ký.	Số Nguyên	Feature
11	**PreferedOrderCat**	Danh mục sản phẩm mà khách hàng ưa thích đặt tháng trước.	Chuỗi	Feature
12	**SatisfactionScore**	Điểm số hài lòng của khách hàng	Số Nguyên	Feature

		(thang điểm 5).		
13	**MaritalStatus**	Tình trạng hôn nhân của khách hàng.	Chuỗi	Feature
14	**NumberOfAddress**	Tổng số lượng địa chỉ mà một khách hàng đăng ký.	Số Nguyên	Feature
15	**Complain**	Lời phàn nàn từ khách hàng trong tháng trước.	Số Nguyên	Feature
16	**OrderAmountHikeFromlastYear**	Phần trăm tăng trưởng đặt hàng trong năm trước.	Số Thực	Feature
17	**CouponUsed**	Tổng số coupon đã sử dụng trong tháng trước.	Số Thực	Feature
18	**OrderCount**	Tổng số đơn hàng được đặt trong tháng trước.	Số Thực	Feature
19	**DaySinceLastOrder**	Ngày mà lần cuối đặt hàng.	Số Thực	Feature
20	**CashbackAmount**	Trung bình tiền trả lại tháng trước (đvt: $).	Số Nguyên	Feature

Bảng 1. Mô tả dữ liệu

CHƯƠNG 3: GIẢI QUYẾT BÀI TOÁN

1. Bài toán 1: Phát hiện điểm đặc thù của các khách hàng trong hệ thống Thương mại điện tử

1.1. Mô tả bài toán

Sử dụng các công cụ thống kê thông dụng như Pivot Table, các hàm của Excel và các dạng lược đồ, biểu đồ để phát hiện, thể hiện các điểm đặc thù của dữ liệu và mối quan hệ giữa chúng.

1.2. Mô tả nguồn dữ liệu và cấu trúc của dữ liệu

Bài toán sử dụng nguồn dữ liệu đã được xử lý ở Chương 2. Sau đây là phần mô tả chi tiết dữ liệu từ bảng mô tả dữ liệu cuối Chương 2:

- *CustomerID:* Dữ liệu thu thập của từng người sẽ được ký hiệu bởi 1 con số nhất định và duy nhất.
- *Churn:* Phân thành 2 trường hợp là không rời bỏ hoặc rời bỏ sàn TMĐT.
- *Tenure:* Dao động từ 0 - 61 tháng gắn bó với dịch vụ của tổ chức.
- *PreferredLoginDevice:* Bao gồm 2 thiết bị đăng nhập của khách hàng (Computer, Mobile Phone/Phone).
- *CityTier:* Cấp thành phố từ 1-3.
  **Chú thích:* Ví dụ cụ thể để hiểu CityTier: Tại Việt Nam, TP.HCM và Hà Nội là 2 thành phố lớn nhất cả nước, được xem là thành phố cấp 1. Các thành phố trực thuộc trung ương là thành phố cấp 2. Các thành phố còn lại là cấp 3.
- *WarehouseToHome:* Từ 5 - 127 (km) khoảng cách từ nhà kho đến nhà của khách hàng.
- *PreferredPaymentMode:* Bao gồm 5 phương thức thanh toán (COD, Credit Card, Debit Card, E-Wallet, UPI)
- *Gender:* Phân thành 2 trường hợp là Nam hoặc Nữ.
- *HourSpendOnApp:* Từ 0 - 5 giờ khách hàng dành ra để lướt app hoặc web.
- *NumberOfDeviceRegistered:* Từ 1-6 tổng số thiết bị mà một khách hàng đăng ký.
- *PreferedOrderCat:* Bao gồm 5 danh mục sản phẩm mà khách hàng ưa thích đặt tháng trước (Fashion, Grocery, Laptop & Accessory, Mobile/Mobile phone, Others).
- *SatisfactionScore:* Từ 1-5 điểm số hài lòng của khách hàng.
- *MaritalStatus:* Gồm 3 loại tình trạng hôn nhân của khách hàng (Divorced, Married, Single).
- *NumberOfAddress:* Từ 1 - 22 tổng số lượng địa chỉ mà một khách hàng đăng ký.
- *Complain:* Phân thành 2 trường hợp là khách hàng có hoặc không phàn nàn trong tháng trước.

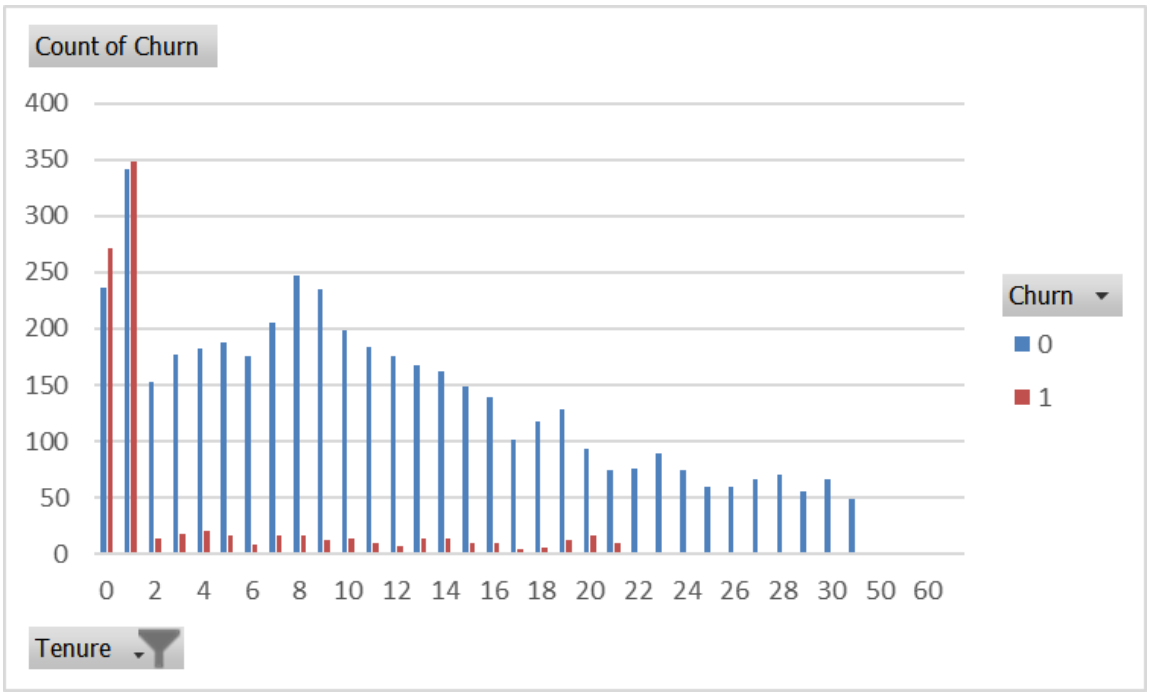
- *OrderAmountHikeFromlastYear:* Từ 11-26% tăng trưởng đặt hàng trong năm trước.
- *CouponUsed:* Từ 0 - 16 tổng số coupon đã sử dụng trong tháng trước.
- *OrderCount:* Từ 1-16 đơn hàng được đặt trong tháng trước.
- *DaySinceLastOrder:* Dao động từ 0-46 ngày, phần lớn từ 0-20 ngày chưa đặt hàng kể từ ngày cuối đặt hàng.
- *CashbackAmount:* Từ 100-325$ trung bình tiền trả lại tháng trước.

1.3. Chọn lọc dữ liệu phân tích

Nhóm sẽ phân tích 10 thuộc tính được chọn lọc dựa trên bảng Rank và phân tích cảm tính, các thuộc tính này được nhóm đánh giá là có ảnh hưởng nhiều đến biến phụ thuộc Churn, từ đó phát hiện, thể hiện các điểm đặc thù của các dữ liệu này và mối quan hệ giữa chúng ảnh hưởng đến quyết định rời đi hay ở lại của khách hàng đối với sàn TMĐT của tổ chức.

1.4. Chạy mô hình và kết quả

1.4.1. Tenure: Thời gian của khách hàng gắn bó với tổ chức

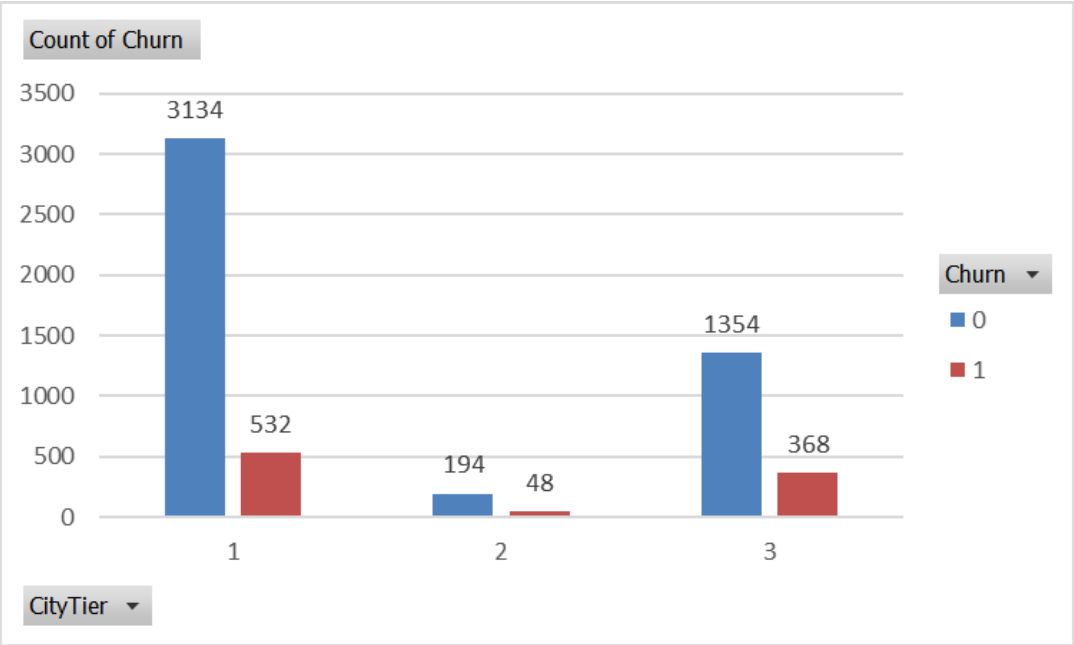


Biểu đồ 1. Tenure

Từ dữ liệu vừa phân tích, cho thấy thời gian khách hàng đã sử dụng web/app thương mại điện tử (TMĐT) phổ biến từ 0-40 tháng. Những khách hàng mới sử dụng dưới 2 tháng, tỷ lệ rời bỏ sẽ rất cao, có thể do sự đa dạng và phổ biến của các sàn thương mại điện tử hiện nay, khách hàng có nhiều sự lựa chọn hơn, nên đối với những khách hàng mới sử dụng, nếu sàn TMĐT chưa đáp ứng mong muốn của họ, họ sẽ rời đi. Ngược

lại, những khách hàng có thói quen sử dụng sàn TMĐT này trên 2 tháng, tỷ lệ rời bỏ sẽ thấp hơn rất nhiều, họ có thể được xem là khách hàng trung thành của tổ chức này.

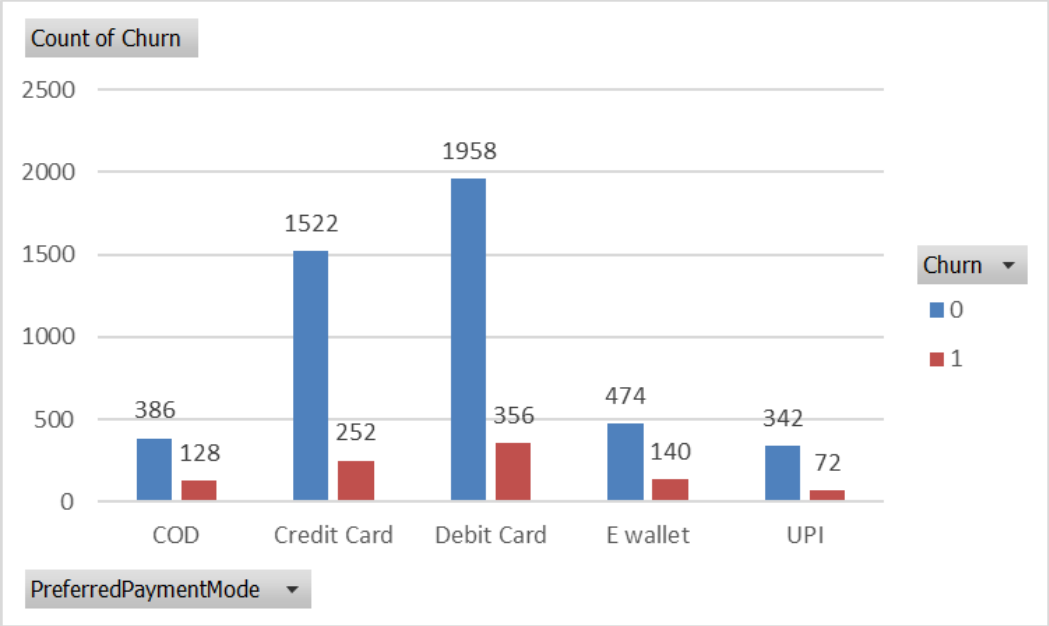
1.4.2. CityTier: Cấp thành phố



Biểu đồ 2. CityTier

Thành phố cấp 1 là thành phố có số lượng khách hàng nhiều nhất, điều này cũng dễ hiểu khi đây là những thành phố phát triển nhất cả nước nên người dân có nhu cầu tiêu dùng cao. Ở thành phố cấp 2, số lượng khách hàng ít nhất trong 3 khu vực, điều này có thể do 1 phần số lượng thành phố cấp 2 của đất nước này chiếm tỉ lệ nhỏ. Thành phố cấp 3 là những thành phố có tỷ lệ khách hàng rời bỏ cao nhất so với tổng lượng khách hàng của khu vực (21.37%)..

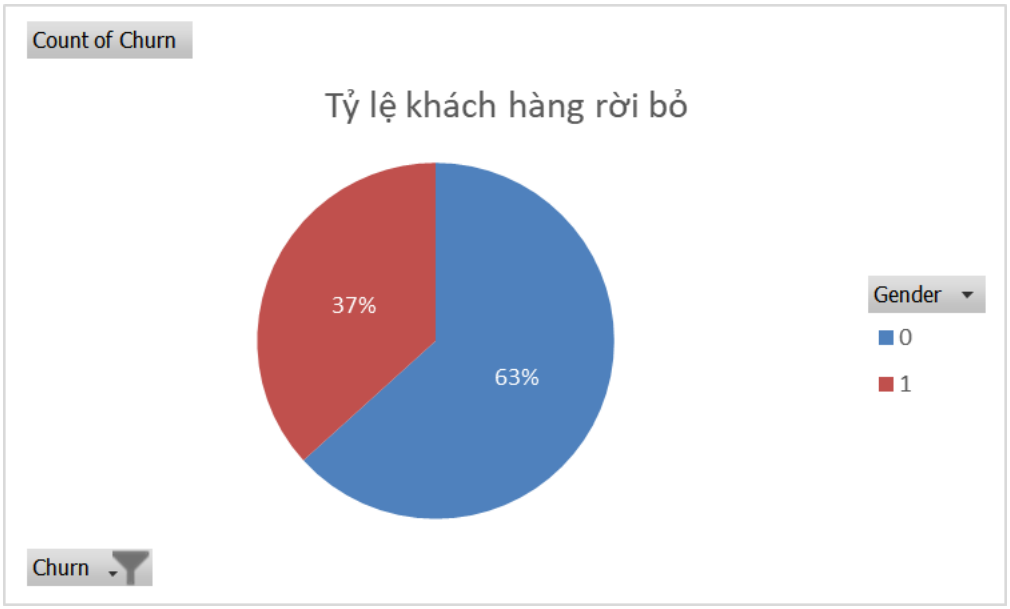
1.4.3. PreferredPaymentMode: Hình thức thanh toán ưa thích của khách hàng.



Biểu đồ 3. PreferredPaymentMode

Hai hình thức thanh toán phổ biến nhất của khách hàng chính là Debit Card và Credit Card, tỷ lệ khách hàng rời đi trên số lượng khách hàng tiếp cận mỗi phương thức của những khách hàng thanh toán bằng hai phương thức này cũng nằm trong top nhỏ nhất (Credit Card là 14.2%, Debit Card là 15.4%), điều này cho thấy đây là hai phương thức được khách hàng xem là tiện lợi. Trong khi đó, tỷ lệ này đối với thanh toán bằng hình thức COD (thanh toán trực tiếp khi nhận hàng) chiếm tỷ lệ cao nhất, có thể do sự bất tiện của hình thức thanh toán này.

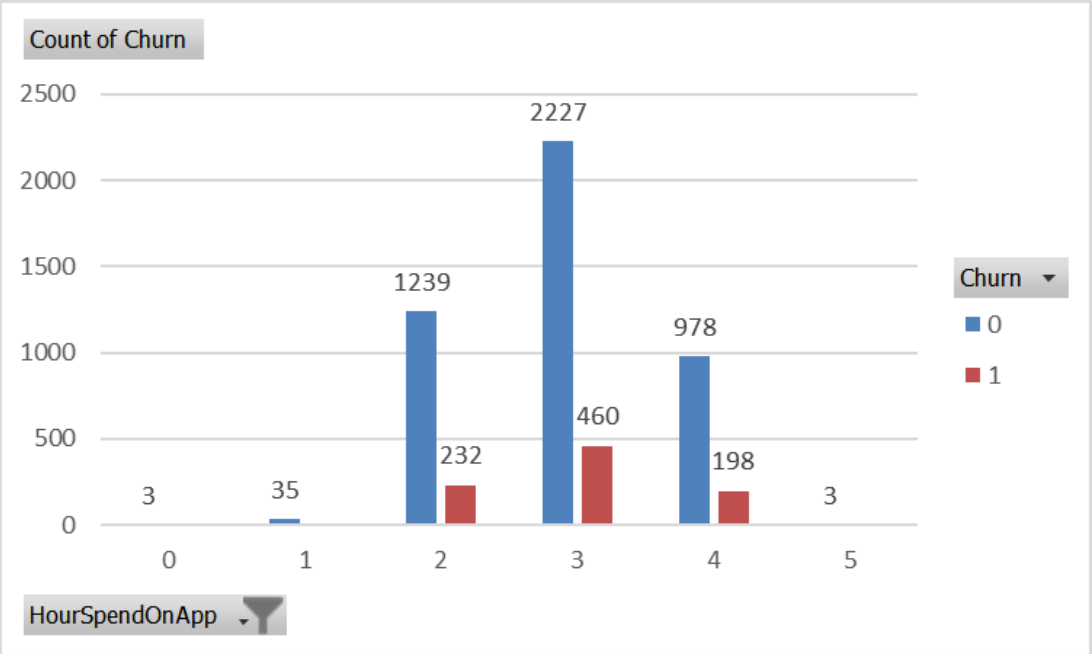
1.4.4. Gender: Giới tính khách hàng



Biểu đồ 4. Gender

Ở biểu đồ này đã thể hiện tỷ lệ khách hàng rời bỏ đi thông qua thuộc tính giới tính. Có thể nhận thấy rằng khách hàng nữ rời đi chiếm tỉ lệ 37% kém hơn gấp 2 lần tỷ lệ khách hàng nam giới rời đi, chạm ở mức 63%.

1.4.5. HourSpendOnApp: Thời gian khách hàng dành ra để lướt app hoặc web

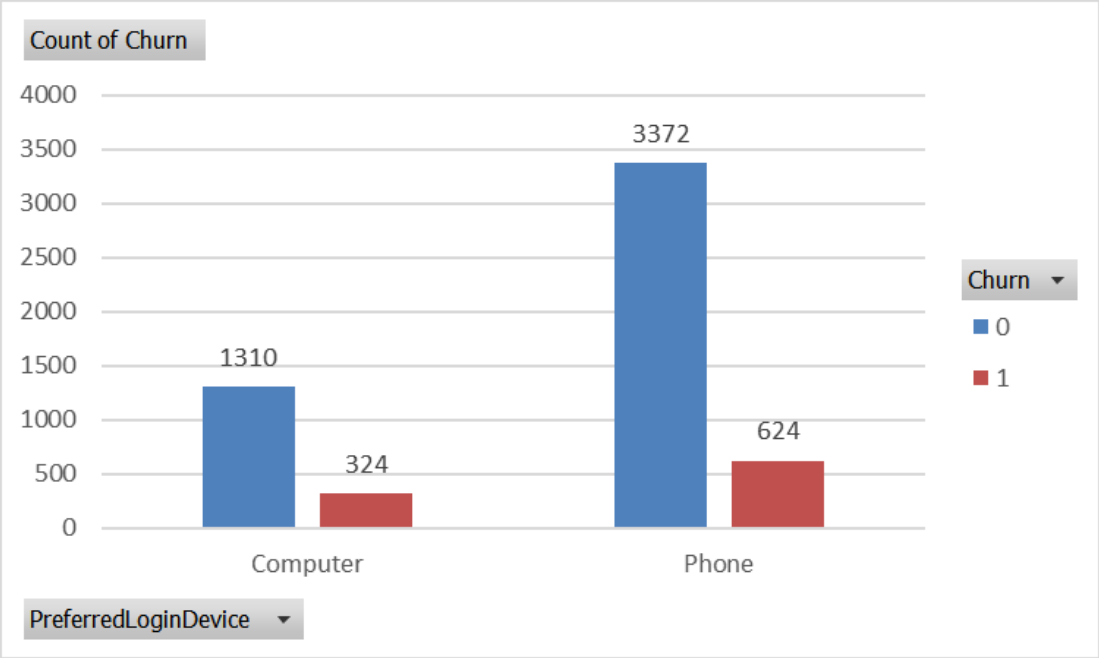


Biểu đồ 5. HourSpendOnApp

Từ biểu đồ cột có thể quan sát rằng phần đông khách hàng có xu hướng dành trung bình từ 2 đến 4 tiếng một ngày cho việc lướt app hoặc website. Lý giải cho hiện tượng này có thể đưa ra một vài lý do như sau: các ứng dụng và website cung cấp cho người dùng nhiều tiện ích và tính năng hữu ích giúp họ tiết kiệm thời gian và công sức. Bên cạnh đó, họ xem việc sử dụng các ứng dụng và trang thương mại điện tử là một công cụ mang đến sự giải trí, một nền tảng mạng xã hội để tương tác với những người khác,...

Xu hướng dành nhiều thời gian hơn nữa cho các nền tảng thương mại điện tử dự báo sẽ còn tiếp tục gia tăng, khi thị trường càng ngày càng năng động và phát triển nhanh chóng. Sự chuyển đổi số từ hình thức mua hàng trực tiếp sang mua hàng trực tuyến chắc chắn sẽ tạo cơ hội cho doanh nghiệp khi biết tạo ra lợi thế cạnh tranh ở thị trường này.

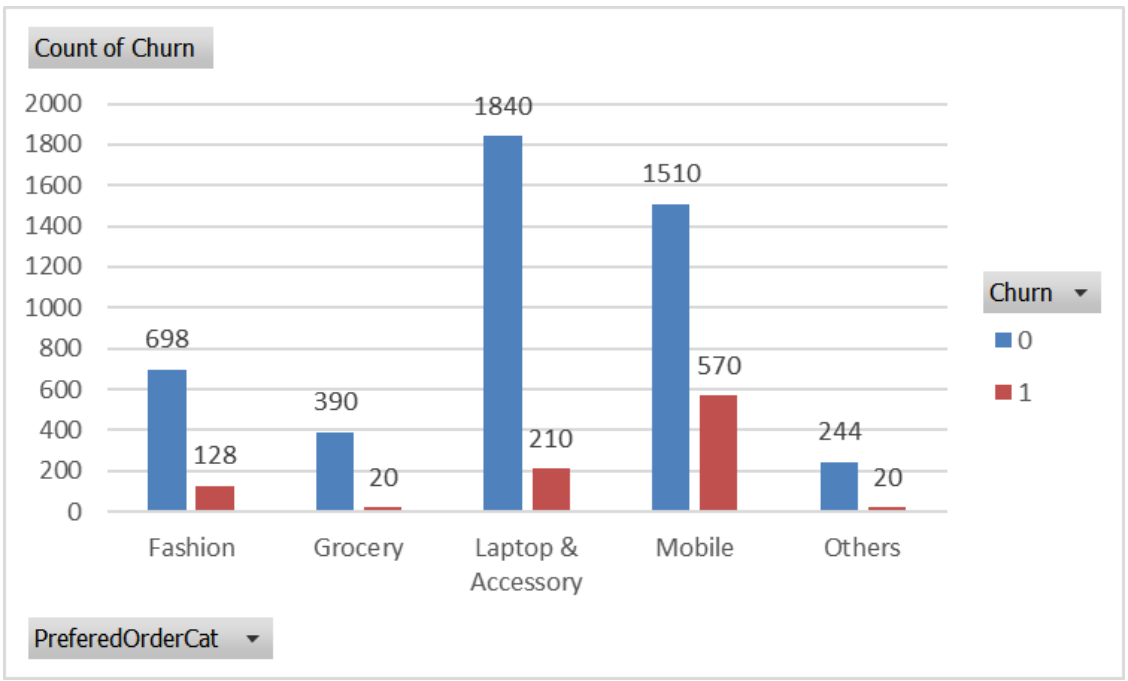
1.4.6.PreferredLoginDevice: Thiết bị đăng nhập ưa thích của khách hàng



Biểu đồ 6. PreferredLoginDevice

Biểu đồ cột cho thấy khách hàng ưa thích sử dụng thiết bị điện thoại di động để đăng nhập và truy cập vào sàn thương mại điện tử cao hơn gần gấp 3 lần so với sử dụng thiết bị máy tính. Bởi vì sự tiện ích, có thể dễ dàng mang theo và truy cập vào mỗi thời gian rảnh tay đã lý giải tại sao điện thoại chính là thiết bị đăng nhập ưa thích của khách hàng.

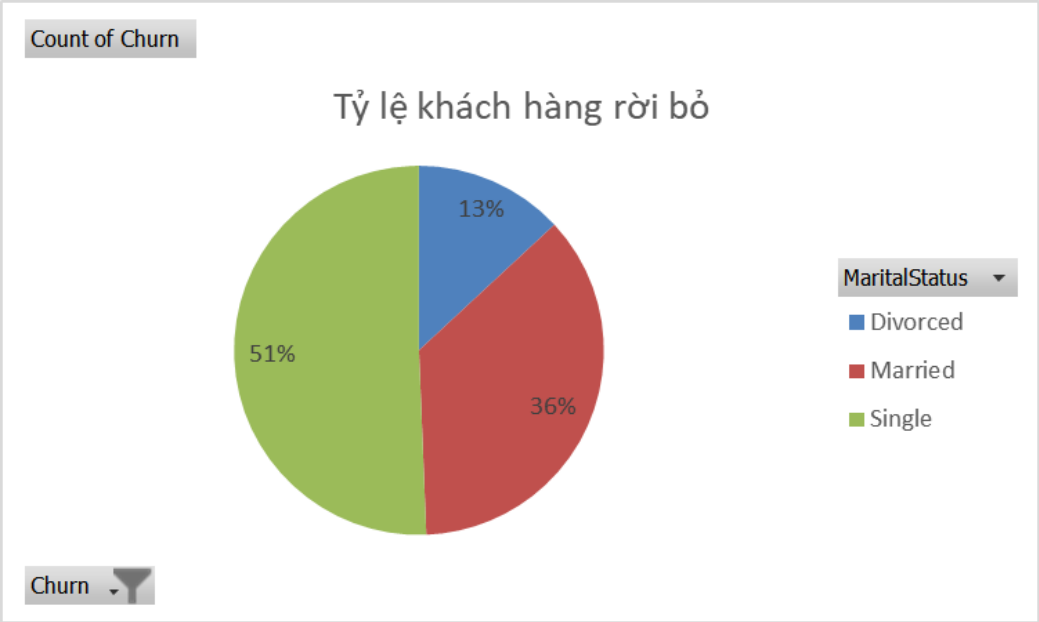
1.4.7. PreferedOrderCat: Danh mục sản phẩm mà khách hàng ưa thích đặt tháng trước



Biểu đồ 7. PreferedOrderCat

Trong số các danh mục sản phẩm ưa thích mà khách hàng đặt từ tháng trước, chiếm tỉ trọng cao nhất ở hạng mục công nghệ là laptop và các phụ kiện cũng như điện thoại di động. Ở hai hạng mục này cũng ghi nhận tỷ lệ ở lại cao nhất so với toàn bộ sản phẩm, có thể bắt nguồn từ nguyên nhân đây là những mặt hàng có giá trị cao, sử dụng lâu dài, khách hàng có xu hướng ủng hộ tiếp tục nền tảng trong lâu dài sau khi nhận được trải nghiệm tốt sau lần mua hàng đầu tiên. Bên cạnh đó, ở hai mặt hàng này cũng ghi nhận tỷ lệ rời đi cao nhất, có thể lý giải từ nguyên do điện thoại, phụ kiện và laptop đều là những sản phẩm không cần thay đổi quá nhiều nên sau khi đã hoàn tất giao dịch, một thời gian dài sau khách hàng mới cần mua lại.

1.4.8. MaritalStatus: Tình trạng hôn nhân của khách hàng

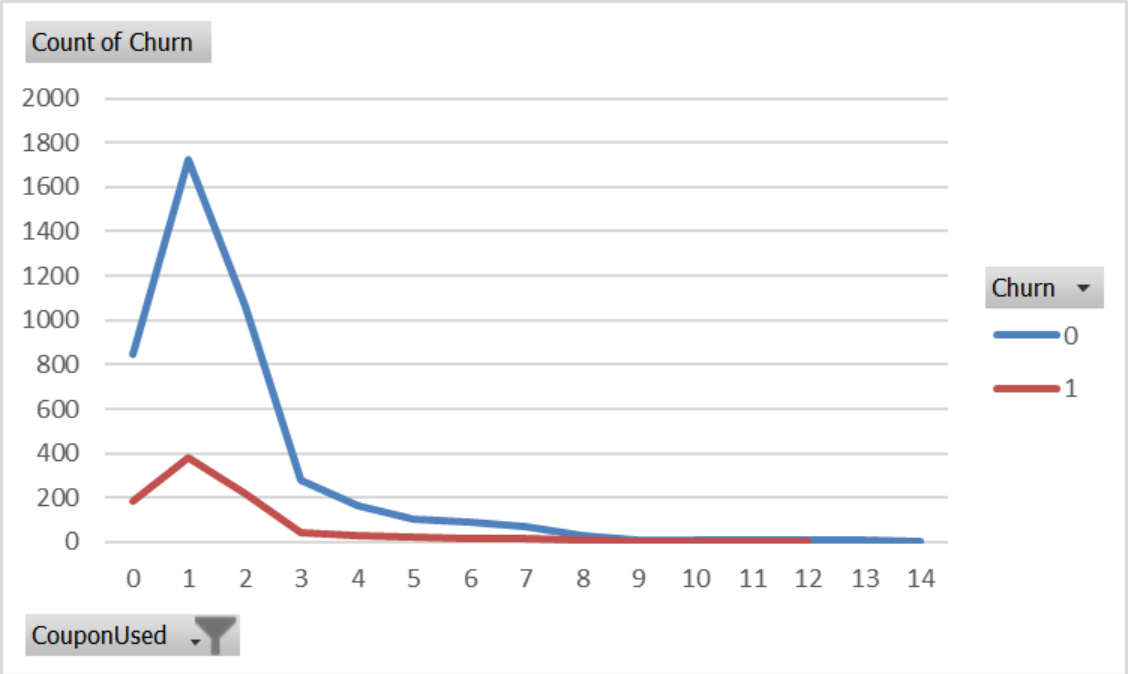


Biểu đồ 8. MaritalStatus

Khách hàng rời đi với tỉ lệ 51% trong biểu đồ tròn thuộc về nhóm khách hàng độc thân. Nhân khẩu học của nhóm khách hàng này thường là người trẻ, tài chính còn chưa ổn định và yêu thích sự tự do. Họ thường không có cam kết với một ai hoặc một nơi cụ thể, vì vậy họ có thể dễ dàng thay đổi và chuyển sang nhà cung cấp dịch vụ khác nếu họ cho rằng giá trị của dịch vụ hiện tại không đáp ứng nhu cầu của họ. Bên cạnh đó, khách hàng độc thân thường không có mối quan hệ sâu sắc với nhà cung cấp dịch vụ do đó họ có thể không cảm thấy có nghĩa vụ gì để ở lại nếu họ không hài lòng với dịch vụ.

Nhóm khách hàng chiếm tỉ lệ 36% chính là những người đã kết hôn, 13% còn lại là tỉ lệ khách hàng rời đi nằm ở nhóm đã ly hôn. Một sự thay đổi lớn về tình trạng mối quan hệ chắc chắn sẽ dẫn đến sự thay đổi về nhu cầu chi tiêu mua sắm ở nhóm đối tượng này.

1.4.9. CouponUsed: Tổng số coupon đã sử dụng trong tháng trước

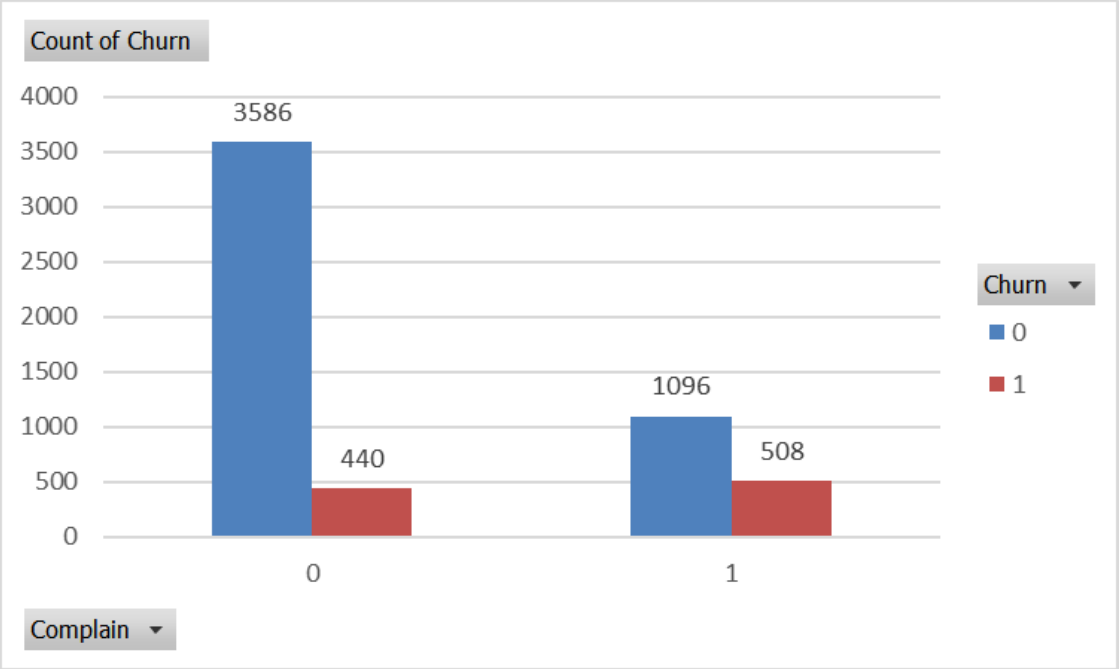


Biểu đồ 9. CouponUsed

Hơn 800 người ở lại sau khi nhận được từ 0 đến 2 coupon, thậm chí đã có gần 1800 người ở lại sau khi nhận được 1 coupon khuyến mãi từ nền tảng. Số lượng người rời đi sau khi nhận từ 0 đến 2 couple chỉ dao động 300 người, thấp hơn rất nhiều so với số lượng ở lại.

Như vậy có thể thấy rằng các chương trình khuyến mãi, coupon tặng kèm vẫn luôn là một trong những cách hiệu quả trong việc giữ chân khách hàng ở lại. Tập trung khai thác ở khía cạnh này có thể giúp doanh nghiệp phát triển doanh thu của mình.

1.4.10. Complain: Lời phàn nàn từ khách hàng trong tháng trước.



Biểu đồ 10. Complain

Từ biểu đồ có thể nhận thấy rằng số lượng khách hàng có trải nghiệm tốt khi sử dụng nền tảng thương mại điện tử là 3586 người và họ chọn ở lại sau lần mua này. Bên cạnh đó, có hơn 1000 người tuy chưa hài lòng về dịch vụ vẫn chọn ở lại và hơn 500 người chọn rời đi. Với tỉ lệ phân nửa như thế này, điều quan trọng là doanh nghiệp phải có hướng chăm sóc, xử lý, cải thiện và bù đắp cho khách hàng sau những trải nghiệm không tốt để có thể giữ chân khách hàng lâu hơn.

1.5. Kết luận về bài toán

Từ kết quả phân tích 10 thuộc tính, cho thấy rằng khách hàng rời bỏ thường có những đặc điểm sau: chỉ mới sử dụng dịch vụ của sàn TMĐT dưới 2 tháng, sinh sống tại những thành phố không quá phát triển (cấp 3), thường sử dụng hình thức thanh toán bằng tiền mặt, họ chủ yếu là nam và còn độc thân, thời gian truy cập trang web không quá nhiều (dưới 2h), sử dụng máy tính để truy cập, mua các loại hàng hoá về thời trang và điện thoại, thường xuyên có lời phàn nàn cho sàn TMĐT,...

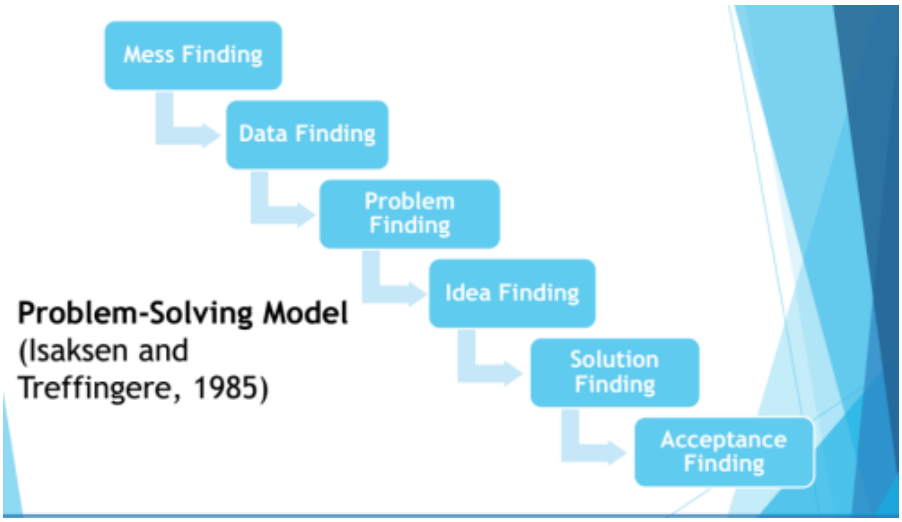
Trong khi đó, những khách hàng có khả năng gắn bó lâu với tổ chức thường có những đặc điểm sau: đã quen sử dụng sàn TMĐT của tổ chức (hơn 2 tháng), sinh sống tại những thành phố lớn phát triển (cấp 1), thường sử dụng các hình thức thành toán như Debit Card và Credit Card, thời gian truy cập trang web dao động từ 2-4h, họ sử dụng điện thoại để truy cập, các loại hàng hoá thường được mua như Computer&Accessory, số lượng coupon được sử dụng trong tháng từ 1-2 cái,...

Qua việc phân tích, nhận ra các đặc điểm đặc thù của khách hàng khi mua sắm tại sàn TMĐT vừa kể trên, tổ chức có thể hiểu rõ hơn về đối tượng khách hàng của

mình, từ đó đưa ra các giải pháp giúp hạn chế việc khách hàng rời bỏ sàn TMĐT, giúp trải nghiệm của khách hàng được tốt hơn, từ đó họ trở thành khách hàng trung thành, gắn bó lâu dài với tổ chức và đem lại một lớn nhuận bền vững cho tổ chức.

1.6. Các kiến thức chuyên ngành đã sử dụng để đánh giá kết quả hay các kiến nghị từ kết quả.

Ứng dụng một phần công việc trong hai bước đầu ở mô hình *Problem - Solving Model* - một mô hình giải quyết vấn đề được học trong môn Phân tích nghiệp vụ, giúp hình thành những kỹ năng ban đầu cho nghề Business Analyst là nghề nghiệp mà phần lớn các bạn trong nhóm đang được học và hướng đến.



Hình 6. Mô hình Problem-Solving

- Ở bước Mess Finding: hiểu được sự phức tạp của tình huống vấn đề. Đối với vấn đề rời bỏ sàn TMĐT, về phía khách hàng, có nhiều yếu tố đang ảnh hưởng tới sự quyết định của khách hàng với việc rời đi hay ở lại (thời gian sử dụng, phương thức thanh toán, khu vực thành phố đang sinh sống, giới tính, tình trạng hôn nhân,...).
- Ở bước Data Finding: Phân tích ý kiến, mối quan tâm, kiến thức và ý tưởng dựa trên dữ liệu. Điều này được thể hiện rõ qua bước chạy mô hình, nhận xét kết quả và suy ra kết luận về bài toán mà nhóm đang thực hiện.

2. Bài toán 2: Dự báo nguy cơ rời bỏ của khách hàng đối với hệ thống Thương mại điện tử và Phân Tích Chuyên Sâu vấn đề hiện tại của hệ thống:

2.1. Mô tả phương pháp

- Phương pháp phân lớp (Classification): Phân lớp dữ liệu là quá trình phân một đối tượng dữ liệu vào một hay nhiều lớp (loại) đã cho trước nhờ một mô hình

phân lớp. Mô hình này đã được xây dựng dựa trên một tập dữ liệu đã được gán nhãn trước đó. Quá trình gán nhãn cho một đối tượng dữ liệu chính là quá trình phân lớp.

- Quá trình phân lớp dữ liệu gồm 3 bước:
  + **Bước 1:** Xây dựng mô hình phân lớp (giai đoạn "huấn luyện").
    - Dữ liệu đầu vào: là dữ liệu mẫu đã được gán nhãn và tiền xử lý.
    - Các thuật toán phân lớp: Cây quyết định, SVM, Hồi quy logistic ...
    - Kết quả của bước này là mô hình phân lớp đã được huấn luyện (trình phân lớp).
  + **Bước 2:** Đánh giá mô hình (kiểm tra tính đúng đắn)
    - Dữ liệu đầu vào: là một tập dữ liệu mẫu khác đã được gắn nhãn và tiền xử lý. Tuy nhiên, lúc đưa vào mô hình phân lớp, ta "lờ đi" thuộc tính đã được gán nhãn.
    - Tính đúng đắn của mô hình sẽ được xác định bằng cách so sánh thuộc tính gắn nhãn của dữ liệu đầu vào và kết quả phân lớp mô hình.
  + **Bước 3:** Phân lớp dữ liệu mới:
    - Dữ liệu đầu vào: là dữ liệu "khuyết" thuộc tính cần dự đoán lớp (nhãn)
    - Mô hình sẽ tự động phân lớp (gán nhãn) cho các đối tượng dữ liệu này dựa vào những gì đã được huấn luyện ở Bước 1.

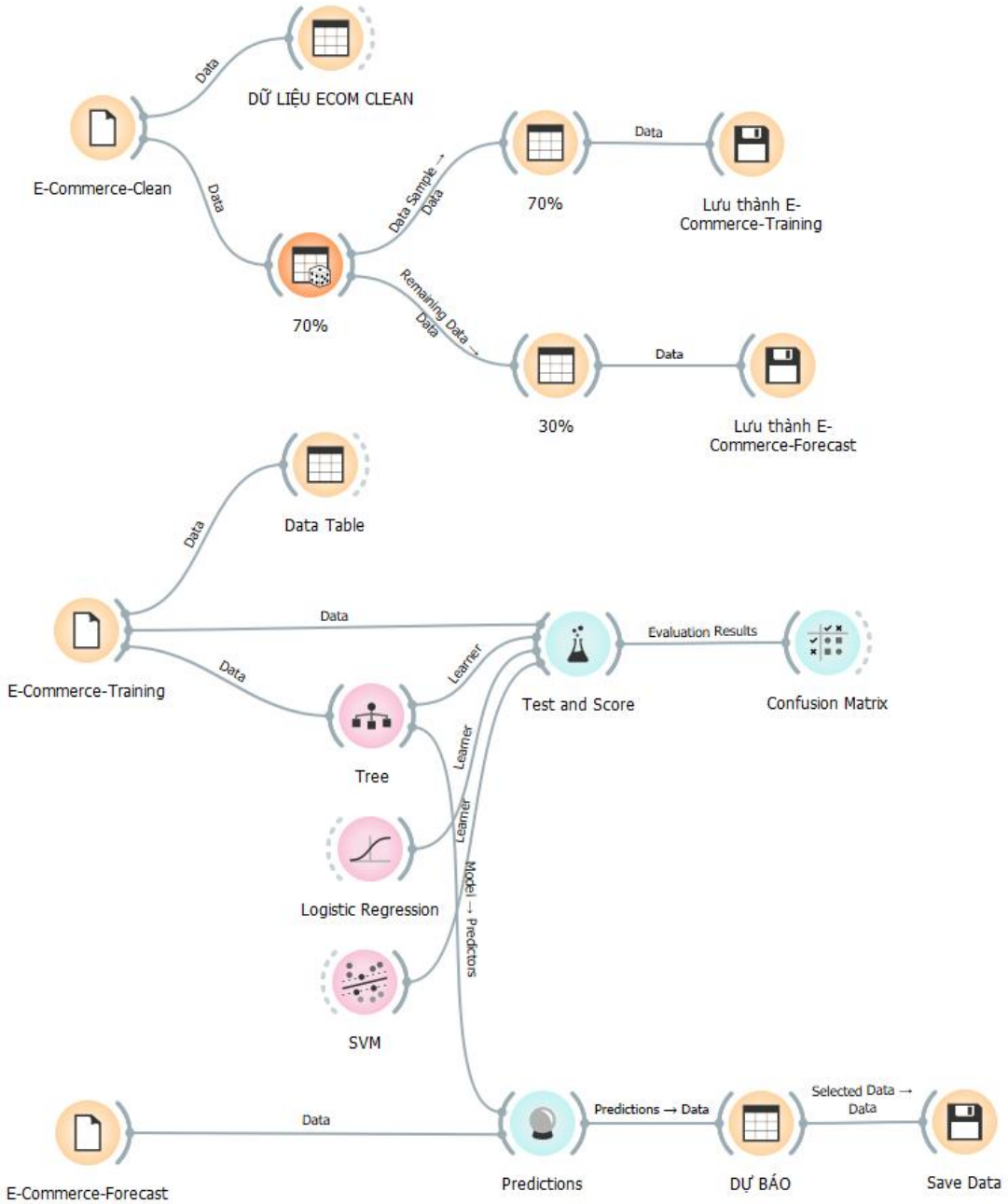
→ Nhiệm vụ của bài toán phân lớp là phân loại đối tượng dữ liệu vào n lớp cho trước. Nếu:

- n = 2: Thuộc bài toán phân lớp nhị phân.
- n > 2: Thuộc bài toán phân lớp đa lớp.

- Các phương pháp phân lớp được sử dụng trong bài:
  + Cây quyết định (Decision Tree): Trong lý thuyết quản trị, cây quyết định là đồ thị các quyết định cùng các kết quả khả dĩ đi kèm nhằm hỗ trợ quá trình ra quyết định. Trong lĩnh vực khai thác dữ liệu, cây quyết định là phương pháp mô tả, phân loại và tổng quát hóa tập dữ liệu cho trước.
  + SVM (Support Vector Machine): là một thuật toán có giám sát, SVM nhận dữ liệu vào, xem chúng như các vector trong không gian và phân loại chúng vào các lớp khác nhau bằng cách xây dựng một siêu phẳng trong không gian nhiều chiều làm mặt phân cách các lớp dữ liệu. Để tối ưu kết quả phân lớp thì phải xác định siêu phẳng (hyperplane) có khoảng cách đến các điểm dữ liệu (margin) của tất cả các lớp xa nhất có thể. SVM có nhiều biến thể để phù hợp với nhiều bài toán phân loại khác nhau.

+ Hồi quy Logistic (Logistic Regression): Là phương pháp nhằm kiểm tra tính hiệu quả của mô hình phân lớp dữ liệu có đặc thù cụ thể, từ đó quyết định có sử dụng mô hình đó hay không. Một mô hình lý tưởng là một mô hình không quá đơn giản, không quá phức tạp và không quá nhạy cảm với nhiễu.

2.2. Quy trình xử lý



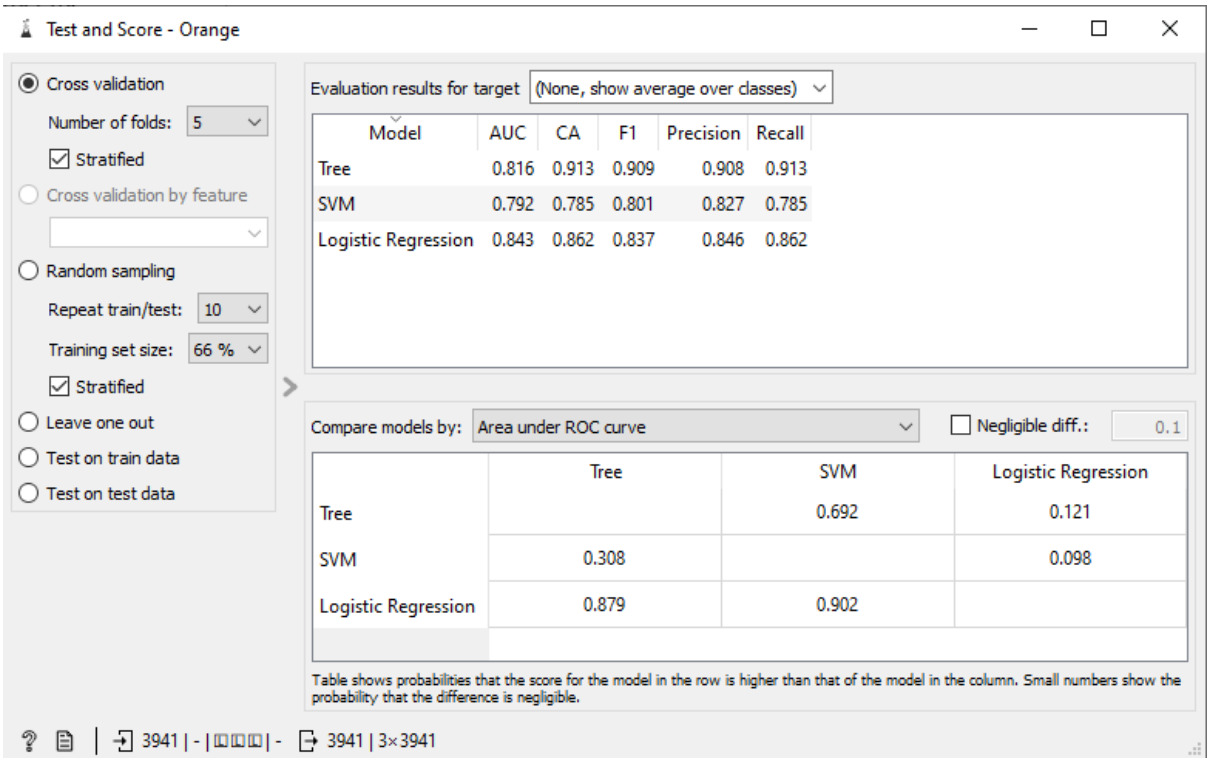
Hình 7. Mô hình Orange xử lý dữ liệu và lựa chọn phương pháp dự báo

- **Bước 1:** Chọn dữ liệu File E-Commerce-Clean.xlsx và chọn cột "Churn" làm Target.

- **Bước 2:** Phân tách dữ liệu: Lọc từ dữ liệu gốc "E-Commerce-Clean.xlsx", nhóm đã sử dụng công cụ Data Sampler tách dữ liệu khảo sát ban đầu thành hai file riêng biệt để thực hiện việc phân lớp dữ liệu như sau:
    - Sử dụng 70% dữ liệu ban đầu để làm dữ liệu mẫu huấn luyện mô hình phân lớp dữ liệu (E-Commerce-Training.xlsx).
    - Sử dụng 30% dữ liệu còn lại để làm dữ liệu dự báo cho nghiên cứu (E-Commerce-Forecast.xlsx).
- **Bước 3:** Dùng 3 phương pháp: Tree, Logistic Regression và SVM tiến hành dự báo rủi ro rời bỏ hệ thống TMĐT và đánh giá độ hiệu quả của các phương pháp.
- **Bước 4:** Lựa chọn phương pháp được đánh giá tốt nhất, dùng phương pháp đó dự báo cho dữ liệu File "E-Commerce-Forecast.xlsx".

2.3. Đánh giá kết quả

Theo Test & Score:



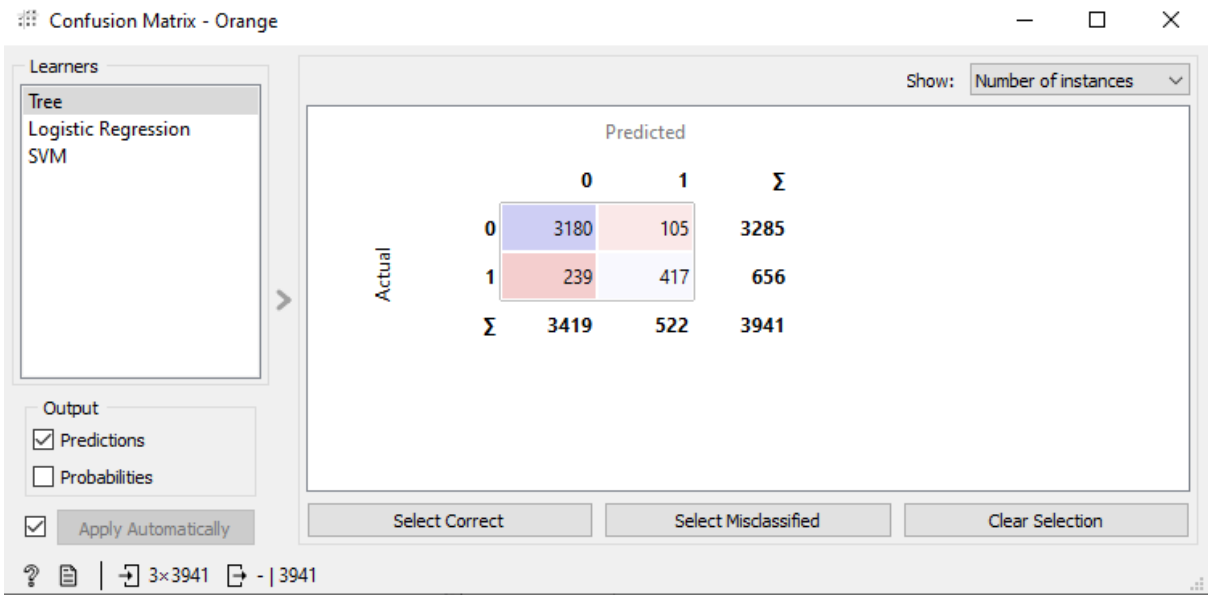
Hình 8. Kết quả dự báo theo Test & Score

$\Rightarrow AUC_{LR} = 0.843 > AUC_{Tree} = 0.816 > AUC_{SVM} = 0.792$

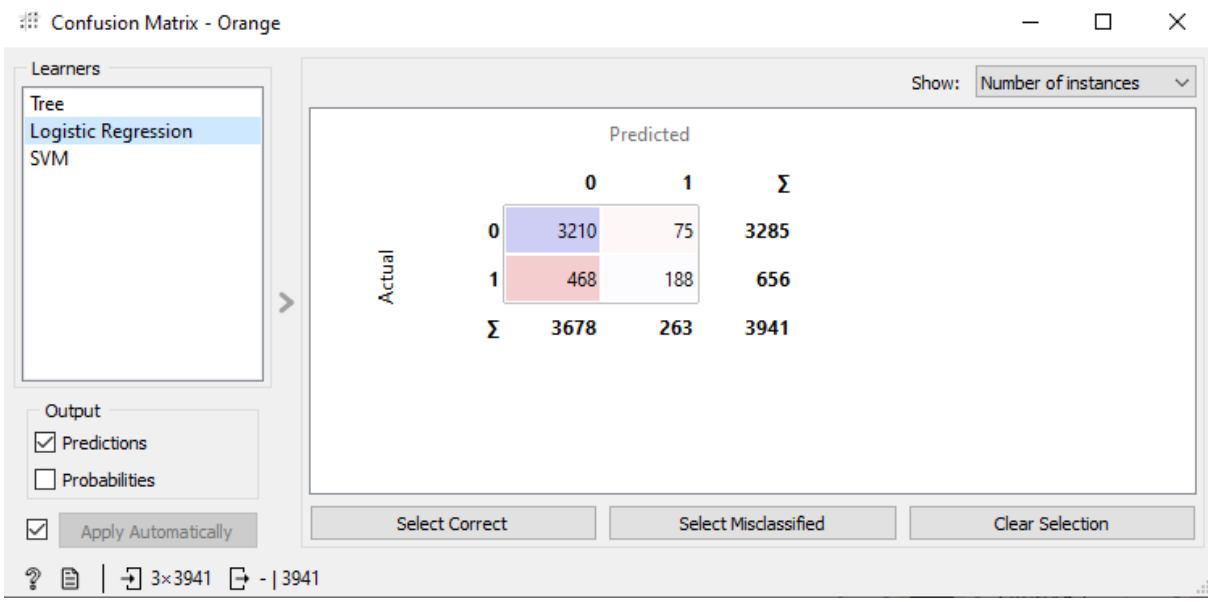
Tuy nhiên: Phương pháp Tree (Cây quyết định) cho ra kết quả Accuracy, F1-score, Precision và Recall là cao nhất trong cả 3 mô hình được sử dụng

→ Nên chọn sử dụng phương pháp Tree (Cây quyết định).

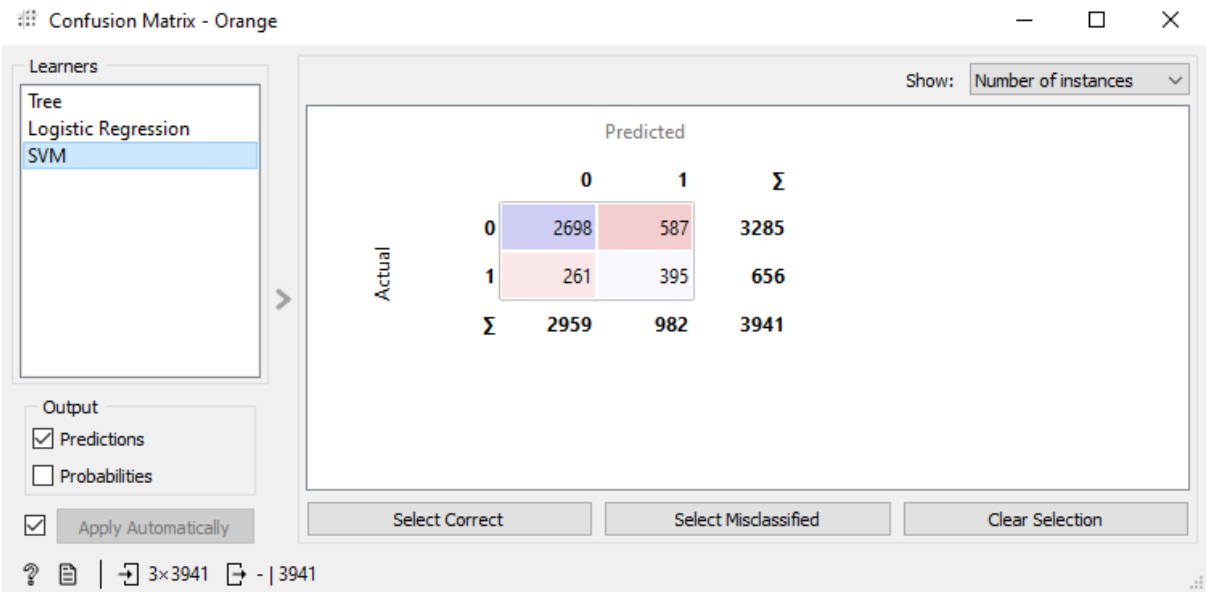
Theo Ma Trận Nhầm Lẫn:



Hình 9. Kết quả dự báo sử dụng phương pháp Tree theo Ma trận nhầm lẫn



Hình 10. Kết quả dự báo sử dụng phương pháp Logistic Regression theo Ma trận nhầm lẫn



Hình 11. Kết quả dự báo sử dụng phương pháp SVM theo Ma trận nhầm lẫn

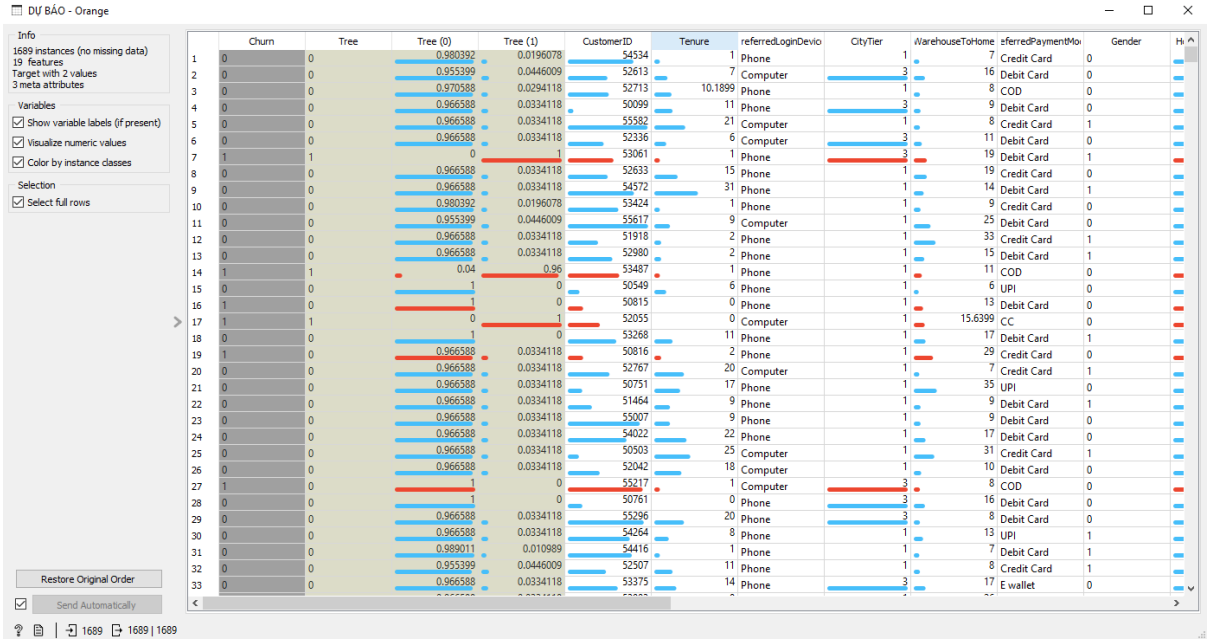
Theo nhóm tự đánh giá ở mô hình này:

- ***Trường hợp 1***: Dự báo khách hàng không rời bỏ hay ngừng sử dụng dịch vụ của công ty (Churn = 0) nhưng Thực tế khách hàng đã rời bỏ hay ngừng sử dụng dịch vụ của công ty (Churn = 1). Bởi vì, khi công ty Dự báo những khách hàng đó không rời đi, thì công ty sẽ không có chính sách, kế hoạch để ưu đãi đặc biệt để giữ chân những khách hàng đó. Điều này sẽ làm mất khách hàng, từ đó gây ra sự sụt giảm về doanh thu và lợi nhuận, ảnh hưởng lớn đến công ty.
- ***Trường hợp 2***: Dự báo khách hàng rời bỏ hay ngừng sử dụng dịch vụ của công ty (Churn = 1) nhưng Thực tế khách hàng không rời bỏ hay ngừng sử dụng dịch vụ của công ty (Churn = 0). Với trường hợp này, công ty sẽ đưa ra ưu đãi đặc biệt cho những khách hàng được công ty Dự báo là rời bỏ, điều này sẽ làm cho công ty tốn một khoảng tiền nhất định để chi cho những khoản phí về ưu đãi đó. Tuy nhiên, khi xem xét ở mặt khác, điều này cũng có thể mang lại lợi ích cho công ty. Bởi vì, khi những khách hàng không có ý định rời bỏ công ty, lại việc được nhận ưu đãi đặc biệt, họ sẽ có cảm nhận và đánh giá tốt hơn về dịch vụ, đồng thời giới thiệu người thân, bạn bè đến công ty sử dụng dịch vụ. Từ đó, công ty sẽ có nhiều khách hàng hơn.

→ Từ đó có thể thấy, trường hợp 1 sẽ gây ra hậu quả nghiêm trọng hơn so với trường hợp 2. Vì vậy, trường hợp 1 sẽ là **Sai lầm loại 2 (Dự báo là không rời bỏ - 0, nhưng Thực tế là rời bỏ - 1).**

⇒ Theo kết quả đánh giá thì mô hình Tree có kết quả sai lầm loại 2 bằng 239 thấp hơn đáng kể so với hai mô hình còn lại (LR = 468, SVM = 261).

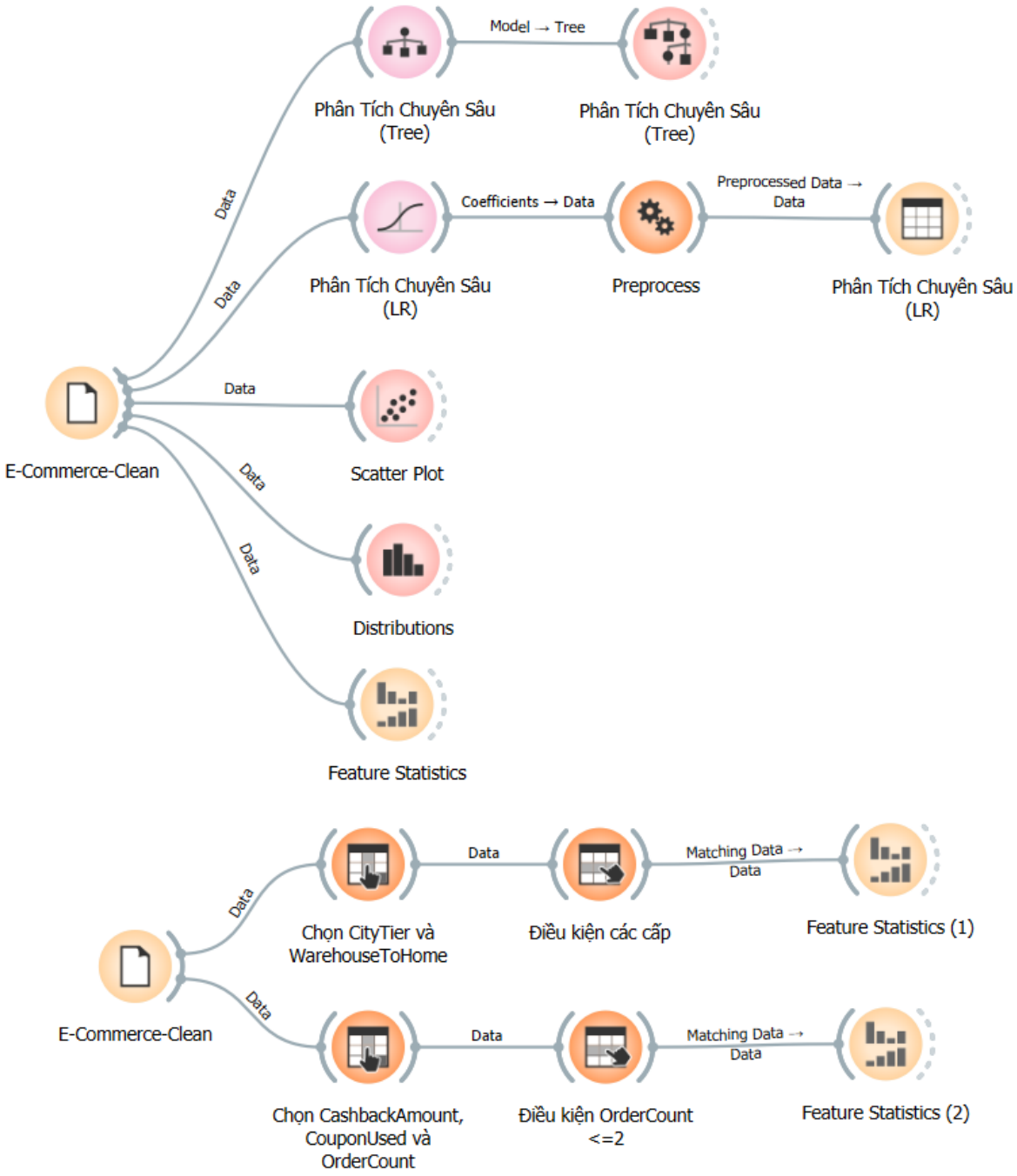
⇒ Từ kết quả trên, nhóm đề xuất công ty nên sử dụng mô hình Tree (Cây quyết định) để dự báo nguy cơ rời bỏ hệ thống Thương mại điện tử hay ngừng sử dụng dịch vụ của khách hàng.



Hình 12. Kết quả Dự Báo khi dùng E-Commerce-Forecast

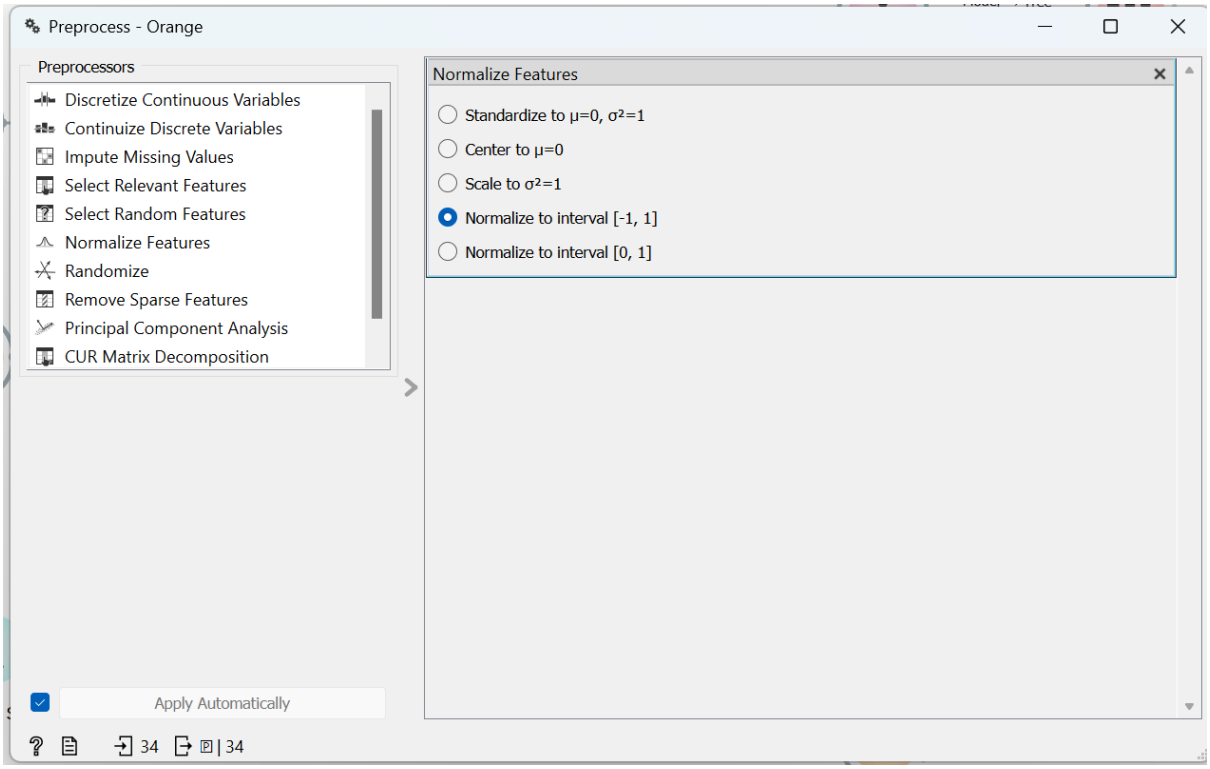
2.4. Phân tích chuyên sâu

Sau khi nhóm đã hoàn thành lựa chọn mô hình để dự báo các khách hàng có nguy cơ rời bỏ Hệ thống Thương mại điện tử, nhóm sẽ tiếp tục tiến hành Phân tích chuyên sâu bằng mô hình Logistic Regression, đồng thời kết hợp với mô hình Tree, các công cụ Distributions (Phân phối giá trị), Feature Statistics (Thống kê mô tả) và Scatter Plot (Đồ thị phân tán) để hiểu rõ hơn về hành vi khách hàng. Từ đó tìm ra các vấn đề đang tồn tại và đề xuất, cung cấp các giải pháp để hạn chế nguy cơ mất khách hàng cho công ty.



Hình 13. Mô hình Phân tích chuyên sâu

Dữ liệu sau khi đưa vào chức năng Logistic Regression, tiến hành xử lý dữ liệu bằng việc chuẩn hóa dữ liệu đó thuộc trong khoảng [-1;1] để có thể thấy được tác động mạnh/ yếu và thuận/nghịch của các thuộc tính trong dữ liệu. Từ đó, dễ dàng thực hiện quá trình Phân tích chuyên sâu.



Hình 14. Chuẩn hóa dữ liệu thuộc khoảng [-1;1] bằng chức năng Preprocess

	name	1
1	intercept	0.57232
2	CustomerID	0.57261
3	Tenure	-1.000
4	PreferredLoginDevice=Computer	0.58723
5	PreferredLoginDevice=Mobile	0.57359
6	PreferredLoginDevice=Phone	0.55709
7	CityTier	0.63795
8	WarehouseToHome	1.000
9	PreferredPaymentMode=COD	0.58561
10	PreferredPaymentMode=Credit Card	0.55553
11	PreferredPaymentMode=Debit Card	0.56104
12	PreferredPaymentMode=E wallet	0.58849
13	PreferredPaymentMode=UPI	0.57283
14	Gender=0	0.58093
15	Gender=1	0.56419
16	HourSpendOnApp	0.59237
17	NumberOfDeviceRegistered	0.66992
18	PreferedOrderCat=Fashion	0.58399
19	PreferedOrderCat=Grocery	0.56979
20	PreferedOrderCat=Laptop & ACredit Cardessory	0.52300
21	PreferedOrderCat=Mobile	0.61154
22	PreferedOrderCat=Others	0.57518
23	SatisfactionScore	0.67706
24	MaritalStatus=Divorced	0.56824
25	MaritalStatus=Married	0.52052
26	MaritalStatus=Single	0.62914
27	NumberOfAddress	0.71850
28	Complain=0	0.49019
29	Complain=1	0.65493
30	OrderAmountHikeFromlastYear	0.54645
31	CouponUsed	0.61861
32	OrderCount	0.62547
33	DaySinceLastOrder	0.31528
34	CashbackAmount	0.56431

Hình 15. Bảng kết quả Logistic Regression sau chuẩn hóa

Dựa trên kết quả của Logistic Regression có thể thấy được các thuộc tính dữ liệu nào sẽ có tác động mạnh yếu và thuận nghịch ra sao đến quyết định rời bỏ Hệ thống Thương Mại Điện Tử (Churn = 1). Từ đó nhóm sẽ đưa ra các đề xuất cho Nhà quản trị để giải quyết tình hình hiện tại và tối ưu hóa lợi nhuận trong tương lai.

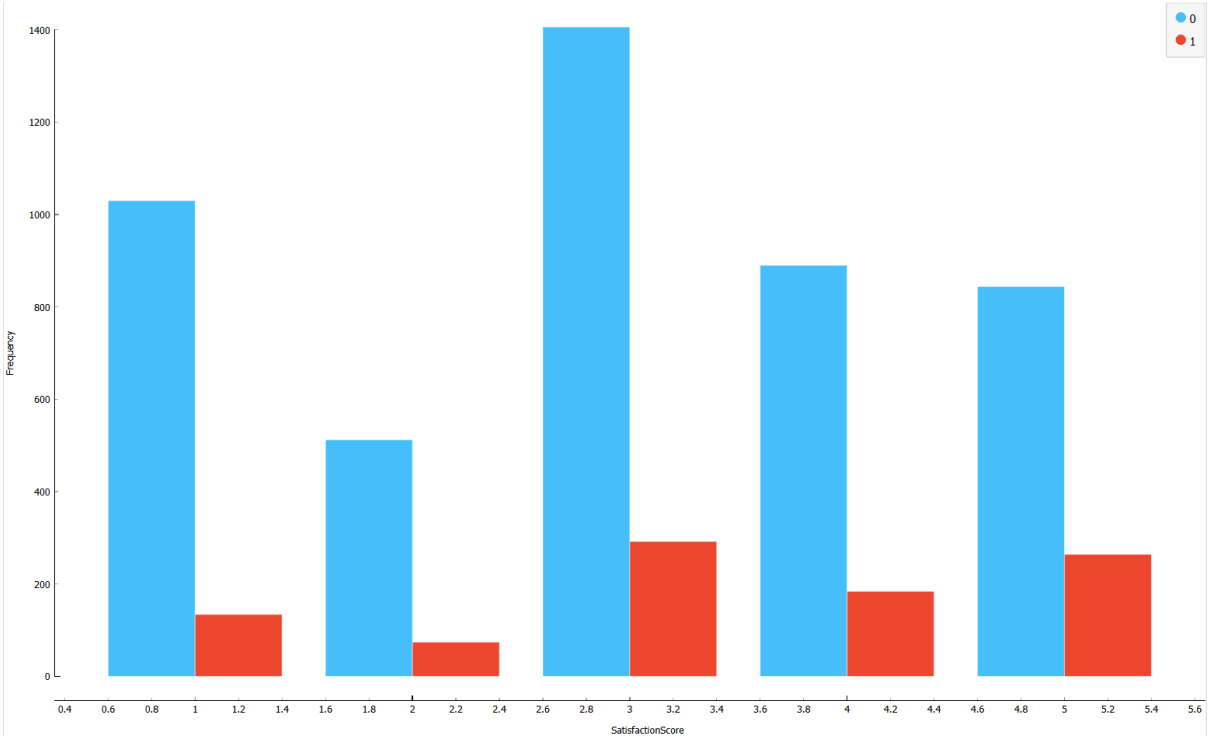
★ Lưu ý:

Các thuộc tính thuộc nhóm PreferredPaymentMode (Hình thức thanh toán ưa thích của khách hàng), PreferredLoginDevice (Thiết bị đăng nhập ưa thích của khách

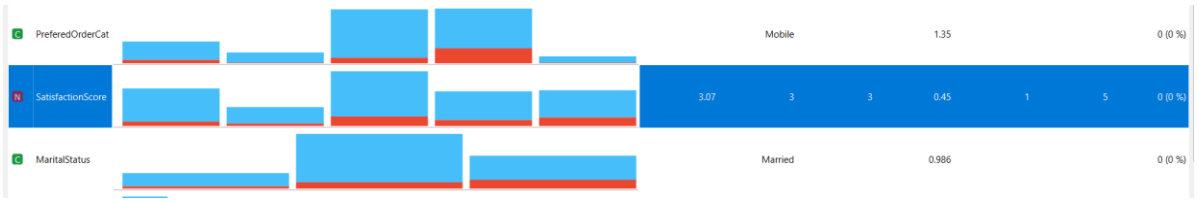
hàng), CustomerID (ID của khách hàng), Gender (Giới tính) sẽ được nhóm quy định là các **Thuộc Tính Ngoại (*)**. Dựa trên phân tích cảm tính, những Thuộc Tính Ngoại là những thuộc tính ảnh hưởng không quá nhiều (hay không ảnh hưởng) đến quyết định rời đi của khách hàng, vì bản chất của các thuộc tính này là nằm ngoài khả năng kiểm soát của Nhà quản trị nên họ sẽ không thể tác động hay làm thay đổi khách hàng để đạt mục tiêu kinh doanh cho công ty. (Ví dụ, thuộc tính PrefferredPaymentMode thể hiện sự ưa thích chủ quan (mức độ tin dùng/ thiên kiến) của khách hàng đối với phương thức thanh toán mà họ sử dụng nên không thể dựa vào thuộc tính đó để đưa ra những thay đổi cho hệ thống. Từ đó, thuộc tính này được xem là Thuộc Tính Ngoại).

★ Phân tích chuyên sâu:

- **Thuộc tính Complain (Lời Phàn Nàn)** thể hiện khả năng rời bỏ hệ thống khá cao và rõ ràng so với các thuộc tính khác. Có thể thấy Complain = 0 (Không có phàn nàn với hệ thống) thì sẽ có tác động tỉ lệ nghịch cao với quyết định rời bỏ đồng nghĩa là các khách hàng càng có ít các đánh giá tiêu cực thì khả năng ở lại và tiếp tục gắn bó càng cao. Ngược lại, Complain = 1 (Có phàn nàn với hệ thống) lại tác động thuận cao với quyết định rời đi, thể hiện rằng khách hàng có càng nhiều đánh giá không tốt thì khả năng cao sẽ rời bỏ hệ thống.
- Một thuộc tính khác cũng cần được chú ý sau khi nhắc đến thuộc tính Complain ở phía trên, là **thuộc tính SatisfactionScore (Điểm số hài lòng của khách hàng**). Khi xét về mặt logic hay độ hiểu thông thường, thì SatisfactionScore sẽ có tác động nghịch cao với quyết định rời đi của khách hàng (điểm hài lòng càng cao, thì khả năng rời bỏ hệ thống càng thấp). Khi được thể hiện bằng bảng kết quả Logistic Regression, thì SatisfactionScore lại tỉ lệ thuận với quyết định rời đi. Vì dữ liệu trong thuộc tính này ban đầu là đầy đủ, không bị thiếu hay lỗi, nên **đây không phải vấn đề xuất phát từ Tiền xử lý Dữ liệu**. Vì vậy, nhóm quyết định sẽ kết hợp phương pháp Tree, công cụ Distributions và công cụ Feature Statistics để phân tích rõ hơn sự bất hợp lí này.

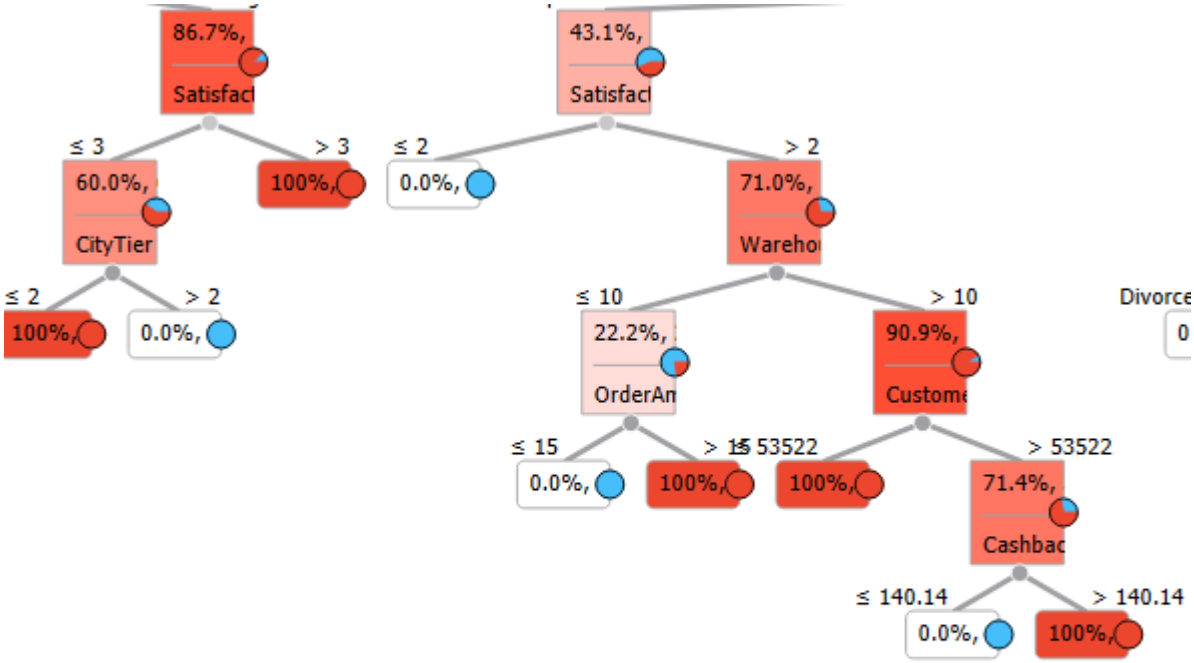


Hình 16. Sơ đồ cột tương quan giữa SatisfactionScore và Churn



Hình 17. Công cụ Feature Statistics cho SatisfactionScore (Colored by Churn)

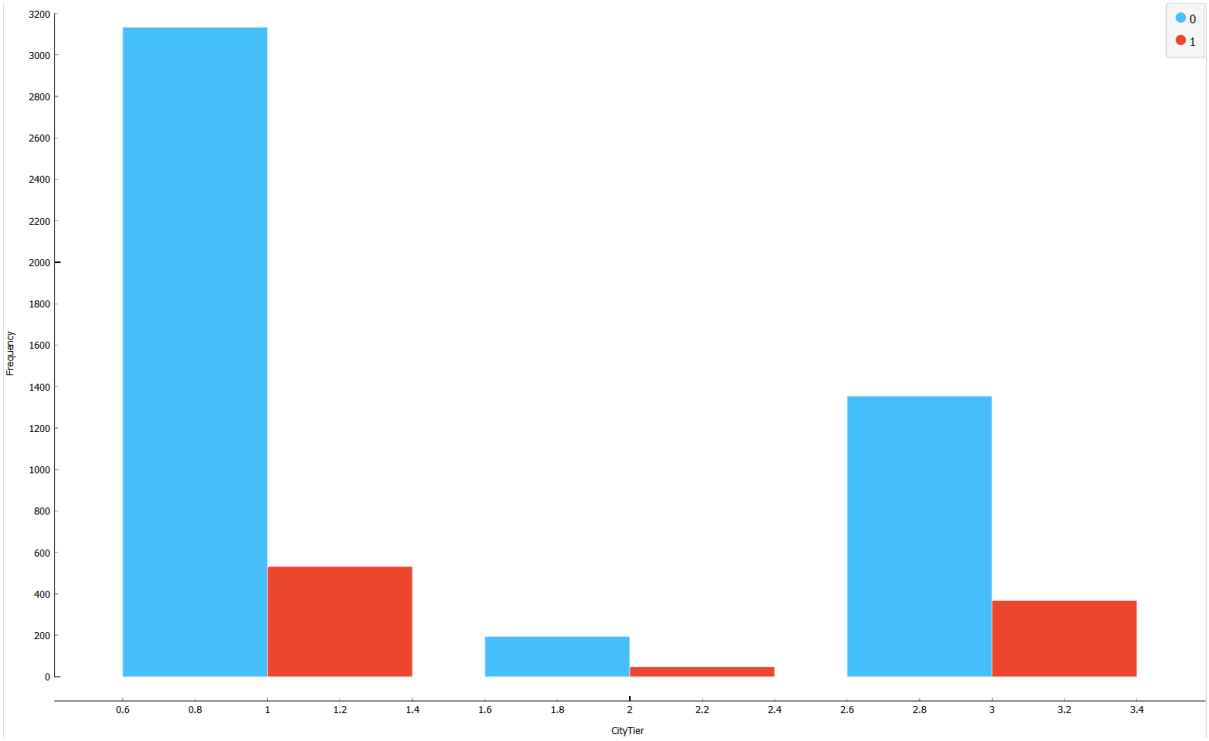
- Qua việc thể hiện SatisfactionScore bằng sơ đồ cột tương quan và công cụ Feature Statistics, nhận thấy điểm trung bình mà khách hàng cho hệ thống là 3 điểm (3.07 điểm) trên thang điểm 5, đồng thời, có sự khác nhau về số lượng người rời đi và ở lại giữa các nhóm điểm. Có 586 người cho 2 điểm (chiếm 10.41% tổng số người cho điểm) nhưng chỉ có 74 người (chiếm 12.63% số người trong nhóm cho 2 điểm hoặc 1.31% trong tất cả các nhóm) là rời bỏ hệ thống, ngược lại ở nhóm 5 điểm - nhóm cho điểm cao nhất - tổng cộng là 1108 người (chiếm 19.68% tổng số người cho điểm) nhưng có đến 264 người (chiếm 23.83% số người trong nhóm 5 điểm hoặc 4.69% trong tất cả các nhóm) là rời bỏ hệ thống. Từ đó, có thể thấy, dù hệ thống nhận được sự đánh giá cao từ khách hàng, nhưng điều đó không giúp Nhà quản trị dự đoán được hành vi của khách hàng là rời đi (hay ở lại) hệ thống. Đồng nghĩa rằng thuộc tính SatisfactionScore phải được phân tích đồng thời với các thuộc tính khác để rõ hơn. Nhóm sẽ tiếp tục dùng Phương pháp Tree để kết hợp phân tích.



Hình 18. Sơ đồ Tree và các thuộc tính gần với SatisfactionScore (Colored by Churn)

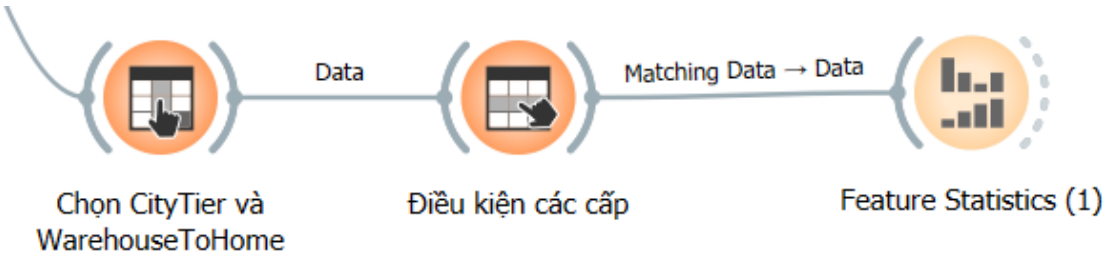
- Dựa trên lý thuyết về Thuộc Tính Ngoại (*), nhóm đã lược bỏ những thuộc tính không ảnh hưởng đến SatisfactionScore và tập trung phân tích những thuộc tính còn lại. Ta có thể đánh giá như sau: SatisfactionScore còn chịu sự ảnh hưởng bởi các yếu tố bên ngoài bên cạnh các yếu tố nội bộ bên trong hệ thống. Ví dụ dễ thấy là sự tương quan giữa CityTier (Cấp Thành phố) và WarehouseToHome (Khoảng cách từ Nhà kho đến Nhà Khách hàng) với SatisfactionScore, sự xuất hiện tương quan giữa 3 thuộc tính này thể hiện rằng khách hàng đang có sự không hài lòng về ***thời gian giao hàng***. Sự trông chờ hàng của khách hàng là có hiện hữu và chính sự trông chờ đó sẽ quyết định sự ở lại hay rời đi của khách hàng. Vì vậy, Nhà quản trị cần phải kết hợp nhiều thuộc tính để đánh giá thay vì chỉ dựa vào số điểm của khách hàng ở thuộc tính SatisfactionScore.
- **Thuộc tính CityTier (Cấp Thành Phố)** và **thuộc tính WarehouseToHome (Khoảng Cách Từ Nhà Kho Đến Nhà Khách Hàng):** nhóm tiến hành phân tích đồng thời 2 nhóm thuộc tính này với nhau để có được kết quả khách quan nhất, bởi 2 nhóm thuộc tính này cùng mang bản chất về vị trí địa lý. Từ Bảng Kết Quả Logistic Regression, có thể thấy được 2 thuộc tính này có tác động thuận khá cao với quyết định rời bỏ hệ thống của khách hàng, tức là Số Cấp Thành Phố

càng tăng thì khách hàng càng có xu hướng rời bỏ hệ thống. Nhóm tiếp tục dùng công cụ Distributions để tìm hiểu và phân tích sâu hơn.

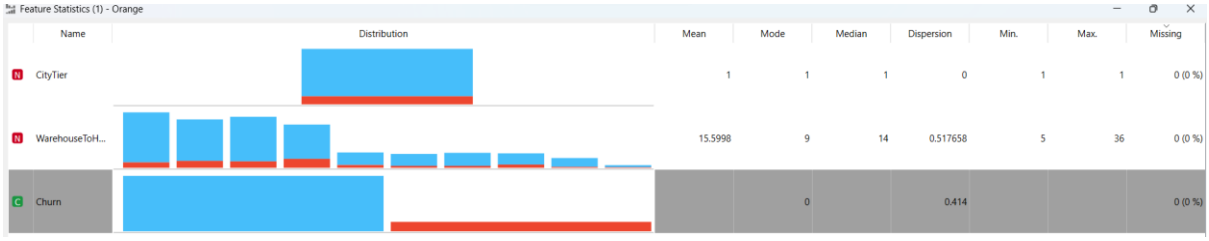


Hình 19. Kết quả Distributions của CityTier (Colored by Churn)

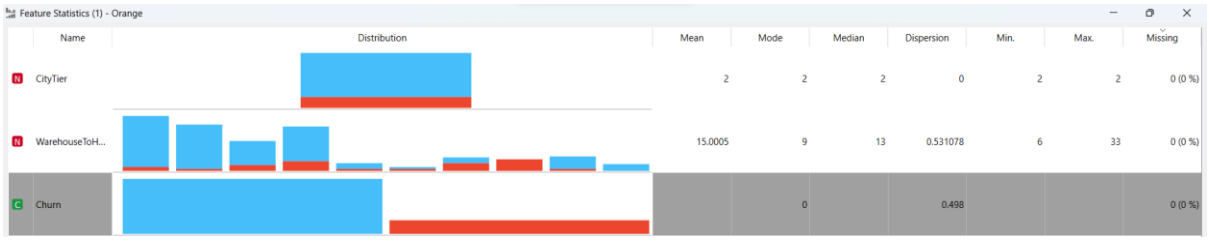
Từ đây có thể thấy rằng khách hàng tập trung nhiều nhất ở Cấp Thành Phố 1 - Cấp Thành Phố lớn nhất, cho thấy hệ thống đang tập trung phát triển ở các Cấp Thành phố lớn, hơn là các Cấp Thành phố khác. Tiếp tục quan sát sự tương quan giữa các cấp CityTier và WarehouseToHome bằng mô hình phân tích sau:



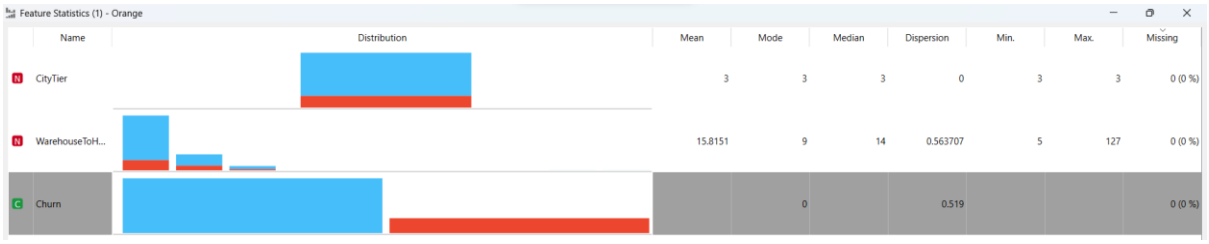
Hình 20. Mô Hình Để Thể Hiện Tương Quan Giữa CityTier và WarehouseToHome



Hình 21. Kết quả Cấp Thành Phố 1



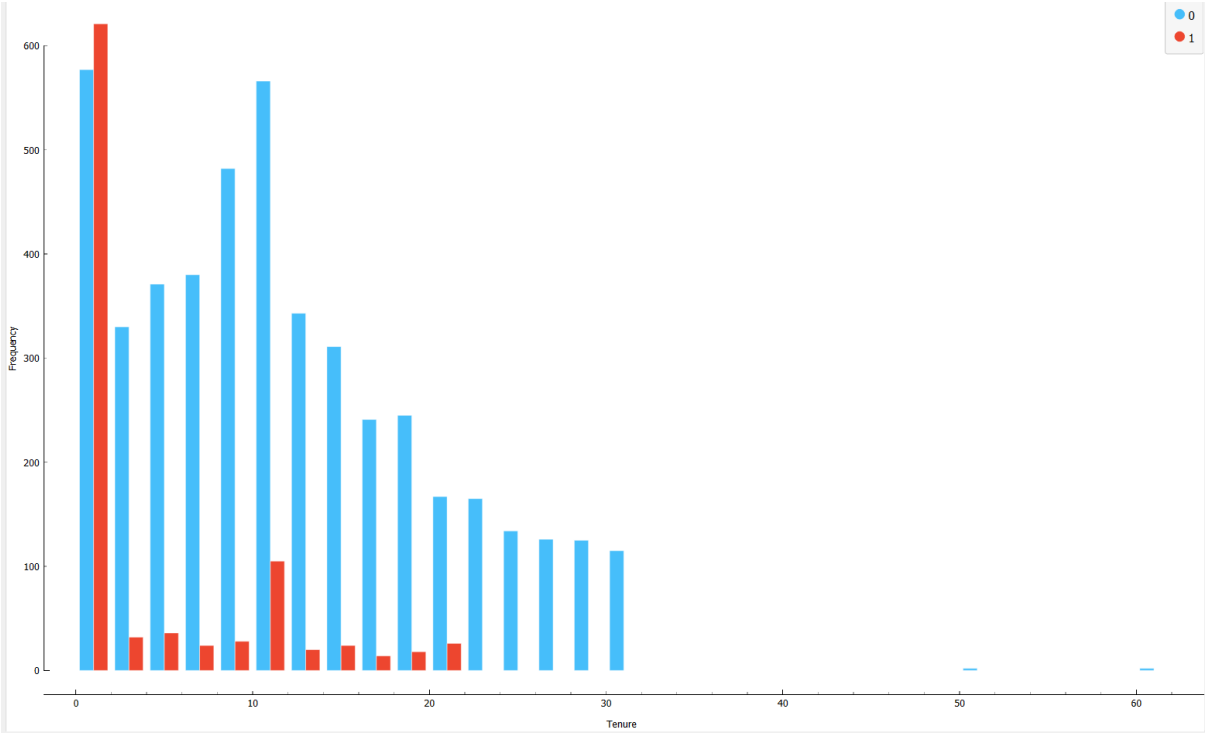
Hình 22. Kết quả Cấp Thành Phố 2



Hình 23. Kết quả Cấp Thành Phố 3
(Colored by Churn)

Việc hệ thống đang tập trung phát triển ở các Cấp Thành phố lớn, hơn là các Cấp Thành phố khác còn được thể hiện thông qua 3 biểu đồ Kết quả các Cấp Thành Phố. 3 biểu đồ cho thấy rằng khoảng cách trung bình từ nhà kho đến nhà khách hàng ở Cấp Thành Phố 3 là cao nhất (mean = 15.815), đồng nghĩa với việc mật độ phân bố các nhà kho ở Thành phố thuộc cấp 3 là rất ít. Kết hợp cùng Bảng Kết Quả Phân Tích Logistic Regression, có thể kết luận rằng: Khoảng cách từ nhà kho đến nhà khách hàng càng lớn (càng xa) thì khách hàng càng có xu hướng rời bỏ hệ thống.

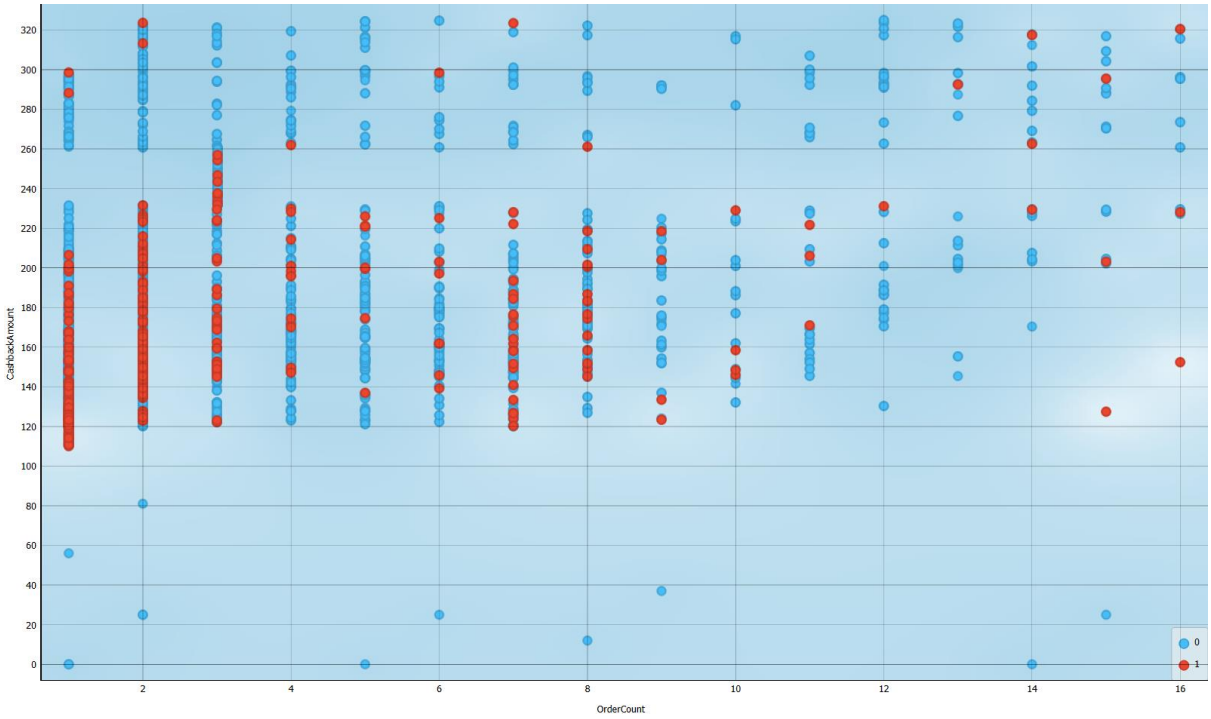
- **Thuộc tính Tenure (Thời gian gắn bó với Hệ thống):** thuộc tính này tác động theo tỉ lệ nghịch cao với quyết định rời bỏ hệ thống. Khi một khách hàng lựa chọn gắn bó với hệ thống từ 22 tháng trở lên họ sẽ có xu hướng gắn bó lâu dài, còn những khách hàng gắn bó từ 2 tháng trở xuống có xu hướng rời đi cao hơn.



Hình 24. Distribution của Tenure
(Colored by Churn)

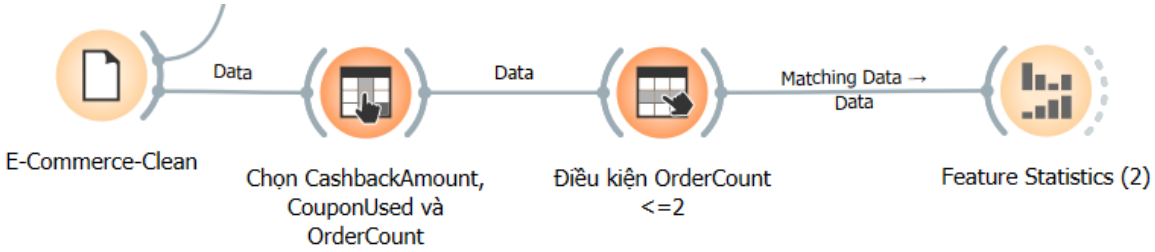
Ngoài ra, còn nhận thấy, số lượng người dùng chỉ gắn bó từ 2 tháng trở xuống là nhiều nhất hay còn có thể nói là nhiều đột biến khi so với các quãng thời gian khác, như vậy trong hệ thống đang tồn tại một vấn đề là: khách hàng dễ gặp khó khăn khi mới sử dụng dịch vụ nên sẽ nhanh chóng rời bỏ hệ thống thương mại điện tử.

- **Thuộc tính CouponUsed (Số phiếu giảm giá khách hàng sử dụng), thuộc tính CashbackAmount (Số tiền trả lại) và thuộc tính OrderCount (Số đơn hàng đã đặt):** Xét Bảng Kết Quả Logistic Regression, sự tác động của các thuộc tính này tới quyết định rời đi của khách hàng là riêng lẻ, độc lập với nhau. Tuy nhiên, theo đánh giá của nhóm khi kết hợp phân tích với Phương pháp Tree, thì các thuộc tính này có mối liên hệ với nhau bởi chúng cùng thuộc trong nhóm Chính Sách Ưu Đãi của công ty dành cho khách hàng. Từ đó, nhóm tiếp tục sử dụng chức năng Scatter Plot để tìm sự tương quan giữa CouponUsed và OrderCount.

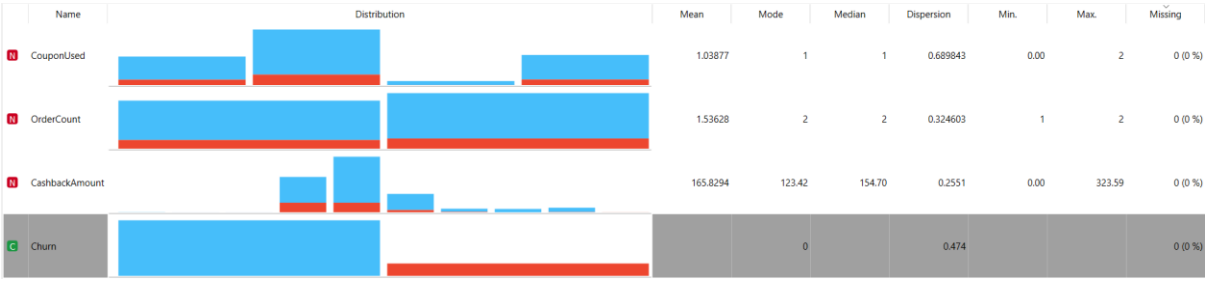


Hình 25. Scatter Plot sự tương quan giữa Order Count và CashbackAmount (Colored by Churn)

Có thể thấy người dùng chỉ có từ 1-2 đơn hàng trên hệ thống tuy nhiên lượng tiền CashbackAmount lại rất cao (trung bình dao động từ 100$ đến 320$). Để chứng minh lượng CashbackAmount này không đến từ CouponUsed, nhóm dùng mô hình sau để thể hiện:



Hình 26. Mô hình phân tích CashbackAmount và Coupon Used



Hình 27. Feature Statistics giữa Coupon Used và CashbackAmount

Có thể thấy rằng, trong nhóm người đặt đơn từ 2 đơn trở xuống, người dùng chỉ sử dụng trung bình 1 Coupon cho mỗi đơn nhưng lượng tiền CashbackAmount lại đạt trung bình 166$. Đồng nghĩa rằng CashbackAmount đến từ hệ thống chứ không từ CouponUsed. Vấn đề ở đây có thể thấy rằng, khách hàng chỉ đặt từ 1-2 đơn và sau khi nhận tiền CashbackAmount thì lại rời bỏ hệ thống ngay, gây thất thoát cho hệ thống.

★ Sau khi đã phân tích và tìm ra các vấn đề mà Nhà Quản Trị có thể thay đổi trong Hệ Thống, nhóm sẽ tiếp tục sử dụng kiến thức chuyên ngành về Business Analysis để đưa ra giải pháp giải quyết vấn đề.

2.5. Kiến nghị cho Nhà Quản Trị bằng kiến thức chuyên ngành

- Kiến thức chuyên ngành sử dụng: với vai trò là một Business Analyst - Người Phân Tích Dữ Liệu Doanh Nghiệp - có trách nhiệm phân tích quá trình kinh doanh của công ty, từ đó xác định vấn đề, đưa ra hướng đi cũng như đề xuất giải pháp thích hợp cho doanh nghiệp. Để có thể đáp ứng được yêu cầu trên, nhóm sẽ tiếp tục sử dụng Mô Hình Giải Quyết Vấn Đề trong nghiệp vụ của Business Analyst:

  MÔ HÌNH GIẢI QUYẾT VẤN ĐỀ
  (PROBLEM-SOLVING MODEL)

  + Mess Finding: Hiểu được sự phức tạp của tình huống vấn đề.
  + Data Finding: Phân tích ý kiến, mối quan tâm, kiến thức và ý tưởng dựa trên dữ liệu.
  + Process Finding: Sử dụng công việc của hai giai đoạn trước để xác định trọng tâm của vấn đề.
  + Idea Finding: Sử dụng các kỹ thuật giải vấn đề, sáng tạo để tìm ra ý tưởng.
  + Solution Finding: Đánh giá các ý tưởng, các giải pháp cho vấn đề.
  + Acceptance Finding: Liên quan đến việc quản lý và thực hiện giải pháp.
- Sau khi dựa vào Mô Hình Giải Quyết Vấn Đề, nhóm đã thực hiện phương pháp Brainstorming giữa các thành viên nhóm để tìm ra vấn đề và từ đó đề xuất các giải pháp (Idea Finding & Solution Finding). Ở bước đưa ra Kiến nghị cho doanh nghiệp hay Nhà quản trị, nhóm sẽ đề xuất giải pháp khả thi cả về mặt ý tưởng lẫn thực tế ứng dụng.
- KIẾN NGHỊ:
  + Vì bản chất của Hệ Thống Thương Mại Điện Tử là một ngành dịch vụ nên Khách Hàng là trọng tâm và là thứ quyết định sự tồn tại của hệ thống. Doanh nghiệp phải lưu tâm đến đánh giá của khách hàng đặc biệt là các đánh giá tiêu cực vì chỉ khi có những đánh giá tiêu cực, hệ thống mới càng ngày hoàn thiện hơn, đáp ứng nhu cầu khách hàng hơn. Đồng thời cũng

phải đặt tầm quan trọng của việc xem xét những đánh giá tiêu cực đó trên "con đường" phát triển và cải thiện hệ thống trong tương lai.

+ Doanh nghiệp cũng sẽ phải phân bổ nguồn lực đa dạng hơn vào thị trường mới, cụ thể ở trường hợp này là vào các Cấp Thành Phố khác thấp hơn để tiếp cận nhiều khách hàng hơn. Song song với đó là xây dựng các nhà kho nhiều hơn và hợp lý hơn vì đây sẽ ảnh hưởng trực tiếp đến tâm lý mua hàng của khách hàng.
+ Như là kết quả nếu như 2 kiến nghị trên được thực hiện một cách hiệu quả, thời gian gắn bó của các khách hàng cũng từ đó tăng lên, đảm bảo một lượng khách hàng thân thuộc, lâu dài cho hệ thống.
+ Doanh nghiệp có thể cần cân nhắc về các Chính Sách Ưu Đãi đang hoạt động trong hệ thống, vì các Chính Sách đó chưa thực sự hoạt động hiệu quả và mang lại giá trị vốn có của nó. Một phương án có thể cân nhắc là dựa vào sự mở rộng (dù hiện tại hay là sẽ có trong tương lai) của hệ thống để thu hút nhiều hơn các bên thứ ba - nơi sẽ cung cấp thêm một phần dịch vụ cho hệ thống, để từ đó cùng với các bên thứ ba đưa ra một Chính Sách Ưu Đãi mới, giảm bớt gánh nặng cho Hệ thống hiện tại.
+ Hệ thống cần thân thiện hơn với người dùng mới, có thể thiết kế giao diện mới đơn giản để người dùng có cái nhìn trực quan, thuận tiện cho việc tự tìm hiểu hệ thống hoặc là cung cấp một phương pháp hỗ trợ hướng dẫn sử dụng cho người mới.

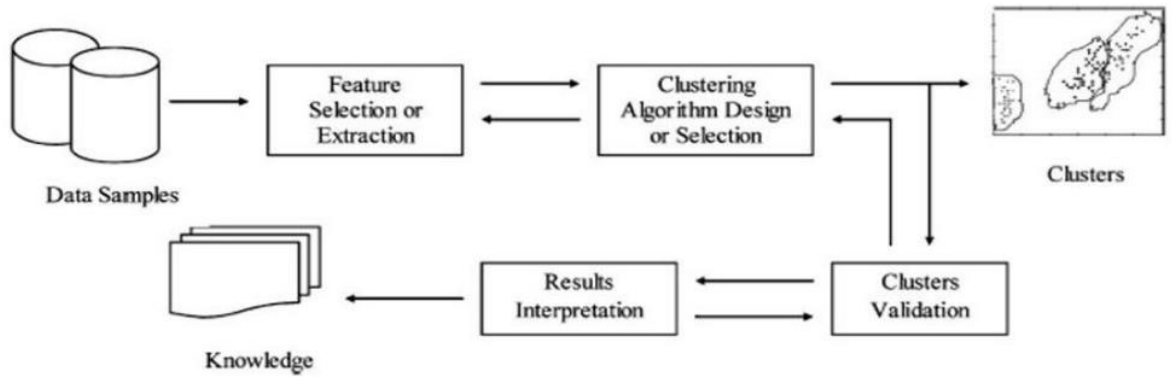
3. Bài toán 3

3.1. Mô tả bài toán

Phân cụm dữ liệu là quá trình gom cụm/nhóm các đối tượng/dữ liệu có đặc điểm tương đồng vào các cụm/nhóm tương ứng. Trong đó:

+ Các đối tượng trong cùng một cụm sẽ có những tính chất tương tự nhau.
+ Các đối tượng thuộc cụm/nhóm khác nhau sẽ có các tính chất khác nhau.

Lưu ý: Dữ liệu của bài toán phân cụm là dữ liệu chưa được gán nhãn. Đây là dữ liệu tự nhiên thường thấy trong thực tế.

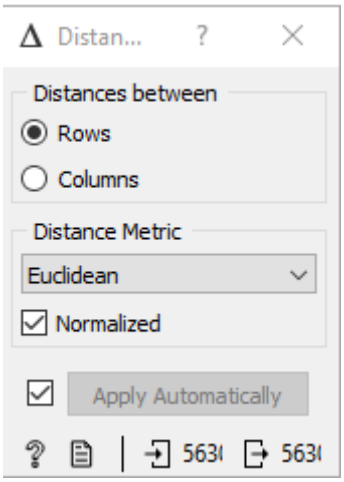


Hình 28. Mô hình phân cụm dữ liệu

3.2. Quy trình xử lý

3.2.1. Phân cụm bằng Hierarchical Clustering

Bước 1: Chọn dữ liệu File E-Commerce-Clean.xlsx và chọn cột "Churn" làm Target.
Bước 2: Chúng ta đo độ phân cụm bằng Distances nhằm tính toán sự tương đồng/sai biệt giữa các đối tượng dữ liệu nhằm phục vụ quá trình phân cụm.

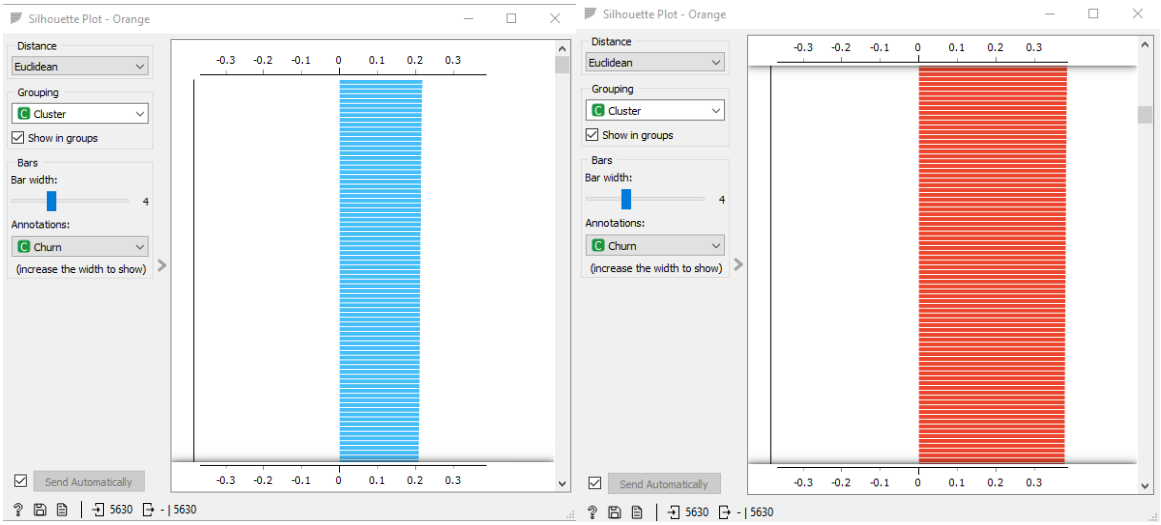


Hình 29. Giao diện Distances

Hình 30. Mô hình phân cụm bằng Hierarchical Clustering

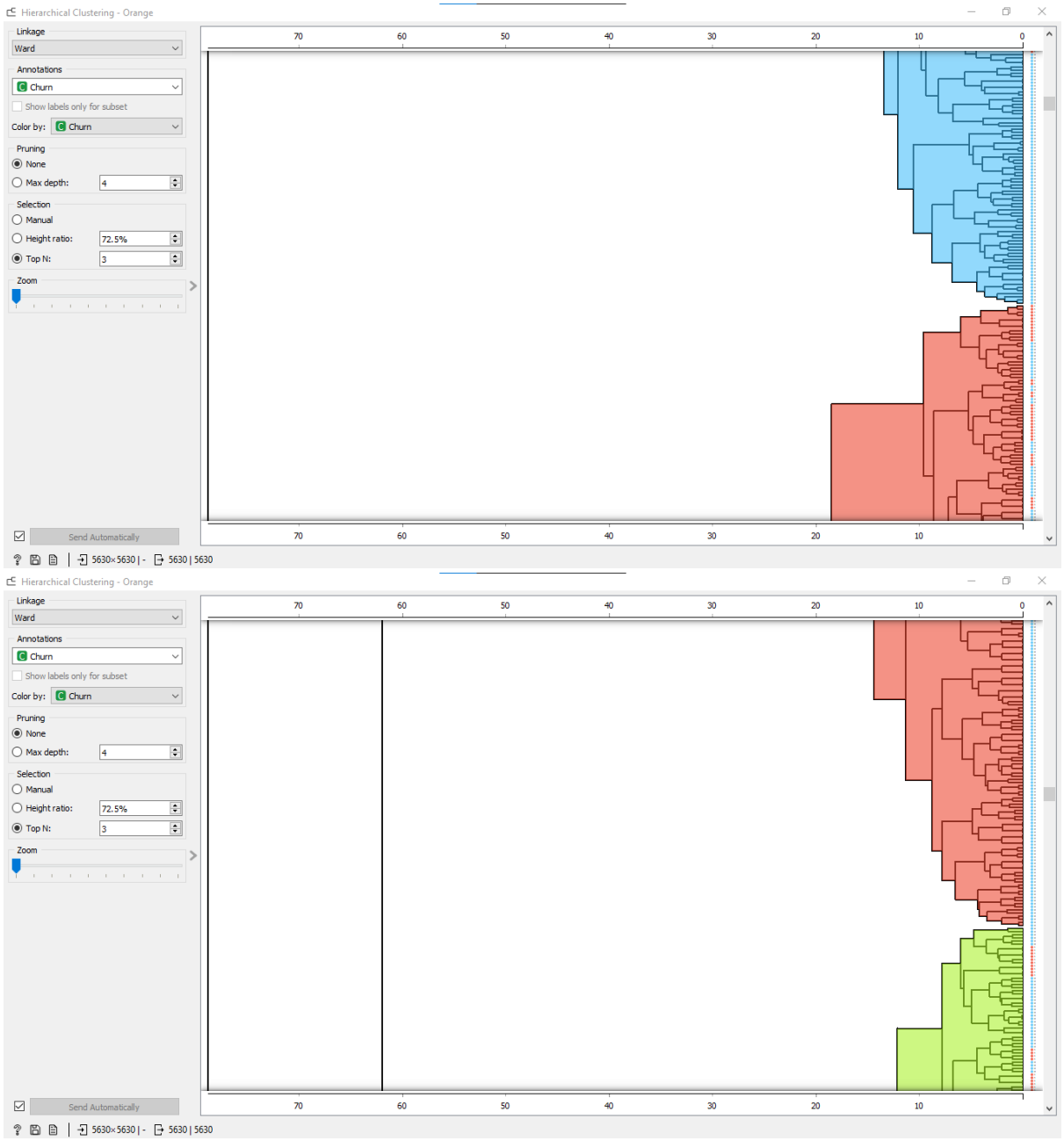
- Hierarchical Clustering: chia dữ liệu thành nhiều cụm khác nhau tùy theo sự điều chỉnh ở mục Selection (chia bằng cách kéo đường gạch đứt nếu chọn Height ratio hoặc tăng giảm giá trị của N để lấy số cụm nếu chọn Top N).
- Silhouette Plot: Để có những đánh giá phân cụm thích hợp, đổ dữ liệu sang công cụ Silhouette Plot, công cụ này giúp đánh giá được độ chính xác của cụm dữ liệu được chia. Khi chỉ số Silhouette càng tiến dần về 1 thì độ chính xác của cụm dữ liệu đã chia càng cao. Chính vì vậy sau khi thực hiện đồng thời 2 bước phân chia dữ liệu và đánh giá, ta có:
    + N=2

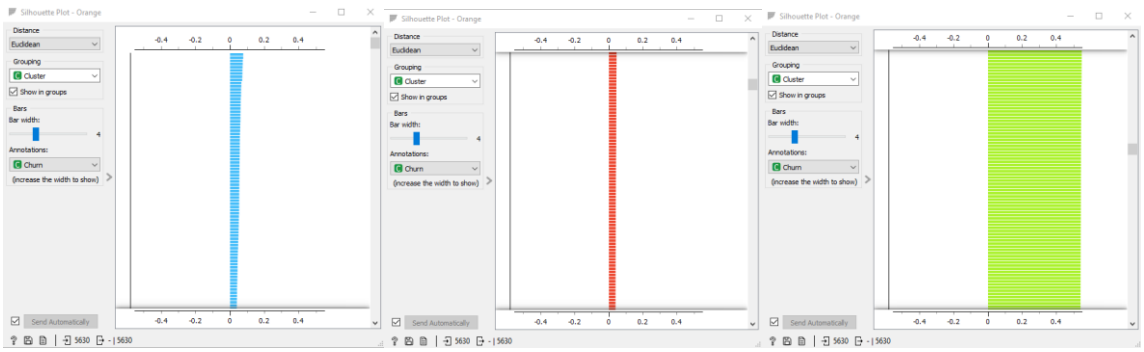
Hình 31. Giao diện Hierarchical Clustering



Hình 32. Giao diện Hierarchical Clustering với số cụm bằng 2

+ N=3

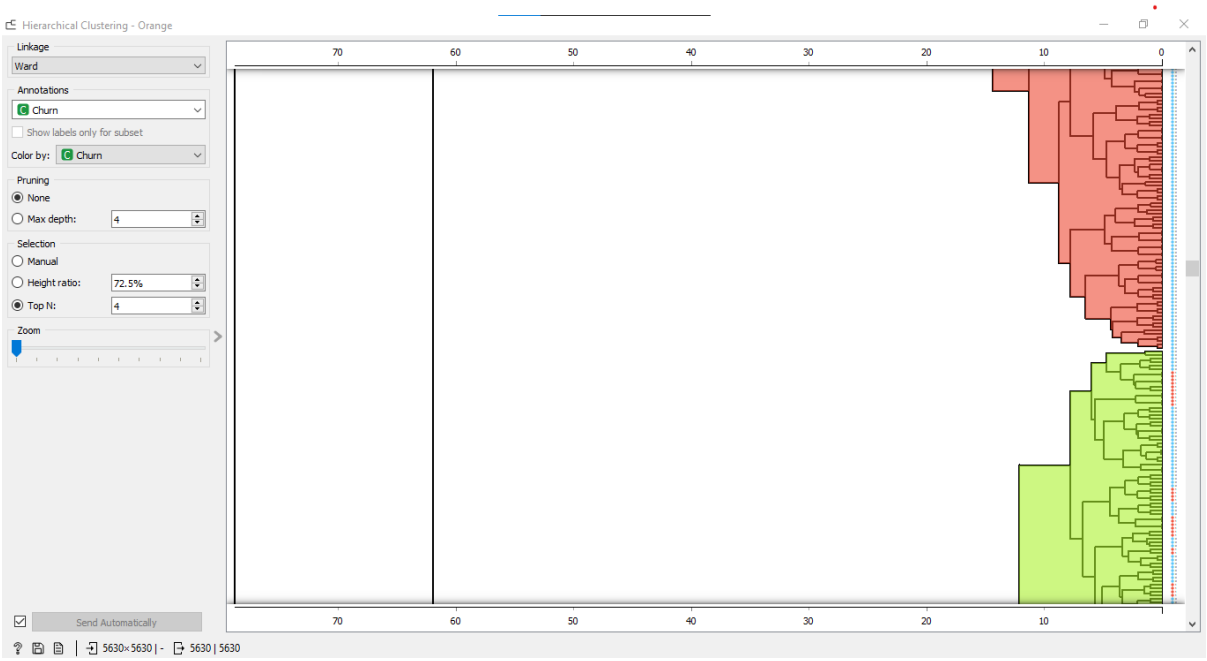


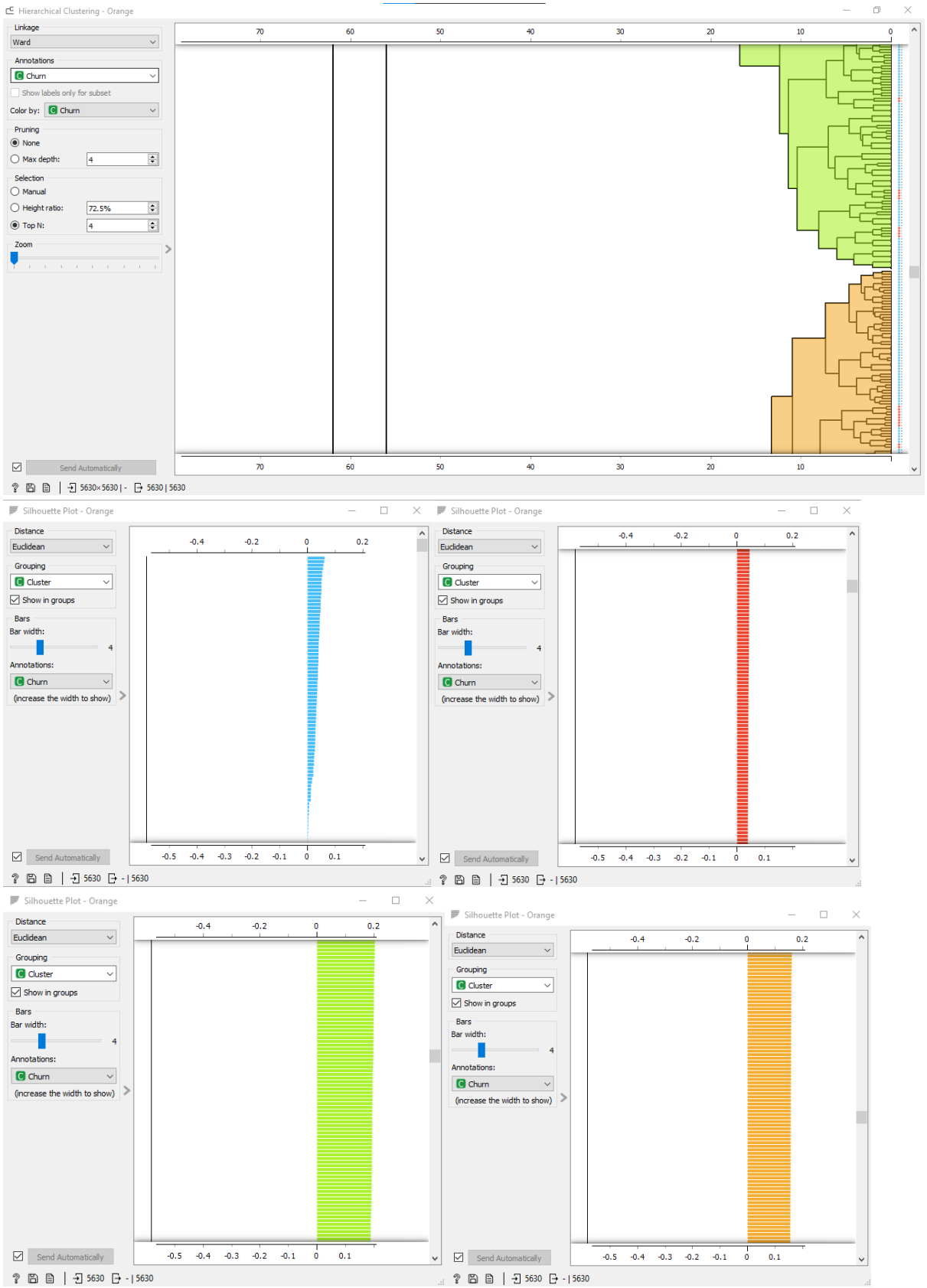


Hình 33. Giao diện Hierarchical Clustering với số cụm bằng 3

+ N=4


Hierarchical Clustering - Orange
Linkage
Ward
Annotations
Churn
Show labels only for subset
Color by:
Churn
Pruning
None
Max depth:
4
Selection
Manual
Height ratio:
72.5%
Top N:
4
Zoom
Send Automatically
5630×5630 | -
5630 | 5630
70
60
50
40
30
20
10
0


Hierarchical Clustering - Orange
Linkage
Ward
Annotations
Churn
Show labels only for subset
Color by:
Churn
Pruning
None
Max depth:
4
Selection
Manual
Height ratio:
72.5%
Top N:
4
Zoom
Send Automatically
5630×5630 | -
5630 | 5630
70
60
50
40
30
20
10
0



Hình 34. Giao diện Hierarchical Clustering với số cụm bằng 4

Ta rút ra bảng kết luận như sau:

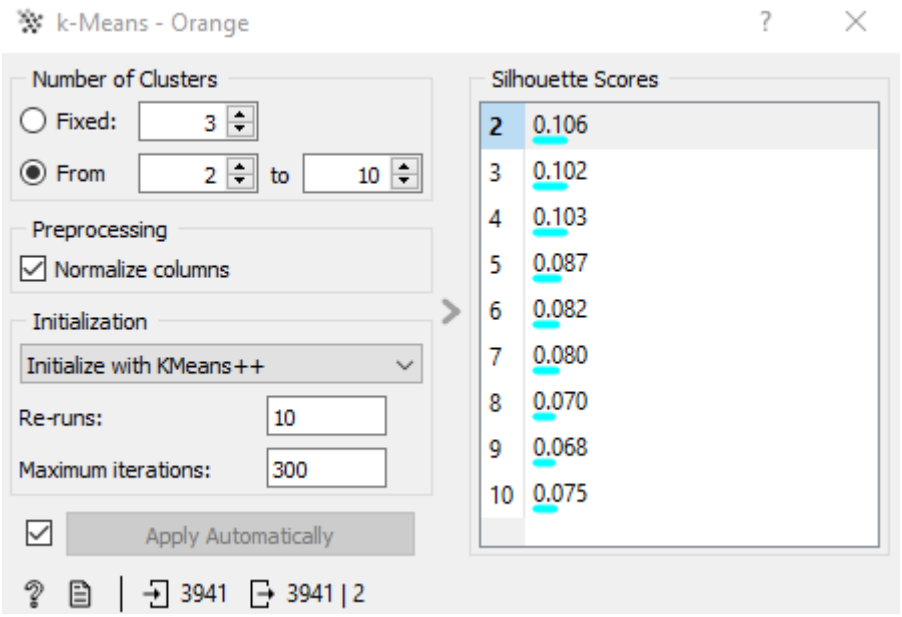
N	Single	Average	Weighted	Complete	Ward
2	0.802	0.802	0.802	0.802	0.217
	0.666	0.666	0.666	0.666	0.384
3	0.802	0.802	0.802	0.802	0.076
	0	0.247	0.247	0.247	0.046
	0.436	0.166	0.166	0.166	0.547
4	0.802	0.802	0.791	0.801	0.061
	0	0.242	0.091	0.239	0.046
	0.860	0.364	0.329	0.291	0.207
	0.436	0.170	0.129	0.171	0.161

Bảng 2. Bảng kết quả tổng hợp chỉ số Silhouette Plot

Bước 3: Từ đây, ta chọn N=2 vì cho chỉ số Silhouette của các cụm là tốt nhất (tiến về phía 1 hơn).Vì vậy, phân thành 2 cụm là tốt nhất.

3.3. Phân cụm bằng K-Means

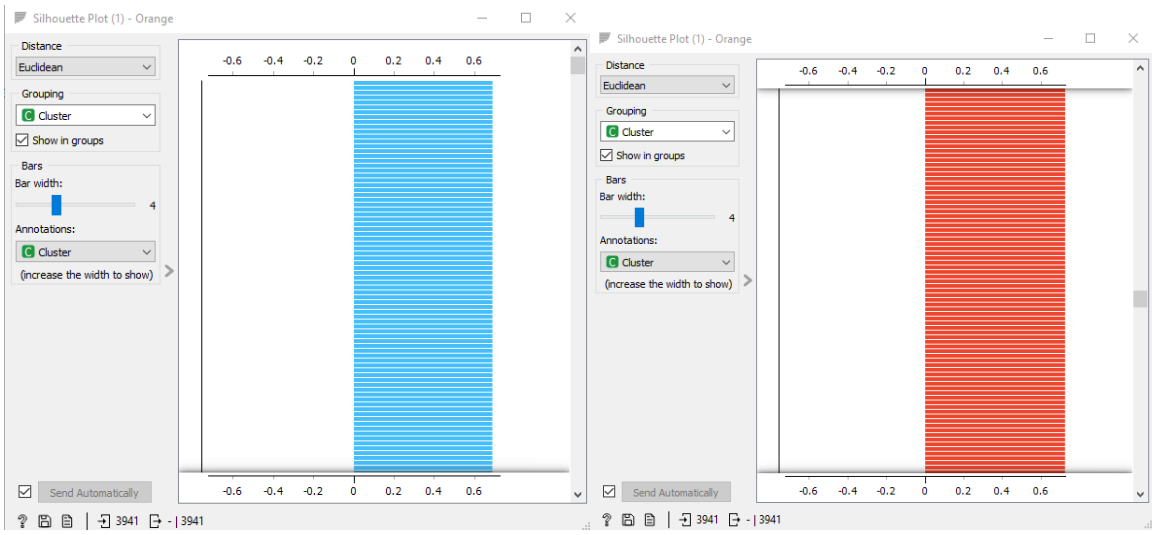
Bước 1: Dùng phương pháp K-Means để phân cụm dữ liệu: Chạy K-Means từ 2 đến 10 cụm, chọn phân thành 2 cụm tương ứng với điểm Silhouette cao nhất là 0.106



Hình 35. Kết quả K-Means

Bước 2: Ta chọn k = 2 vì có chỉ số Silhouette Scores cao nhất và bằng 0.106 ta thấy khi số cụm lớn hơn 2 thì các chỉ số Silhouette Scores càng bé dần. Vì vậy, phân thành 2 cụm là tốt nhất.

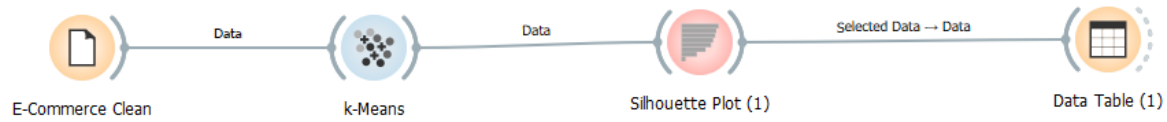
Bước 3: Chọn Silhouette Plot để đánh giá độ chính xác của cụm dữ liệu được chia.



Hình 36. Kết quả Silhouette Plot

	Churn	Cluster	Silhouette	Silhouette (Cluster)
1	0	C2	0.55147	0.72344
2	0	C1	0.524046	0.688247

Hình 37. Chỉ số Silhouette cao nhất của 2 cụm (K-Means)



Hình 38. Mô hình phân cụm bằng K-Means

=> Qua 2 phương pháp Hierarchical Clustering và K - Means, ta có thể thấy, phân thành 2 cụm là phương pháp tối ưu.

3.4. Đánh giá kết quả

3.4.1. Đánh giá các phương pháp phân cụm

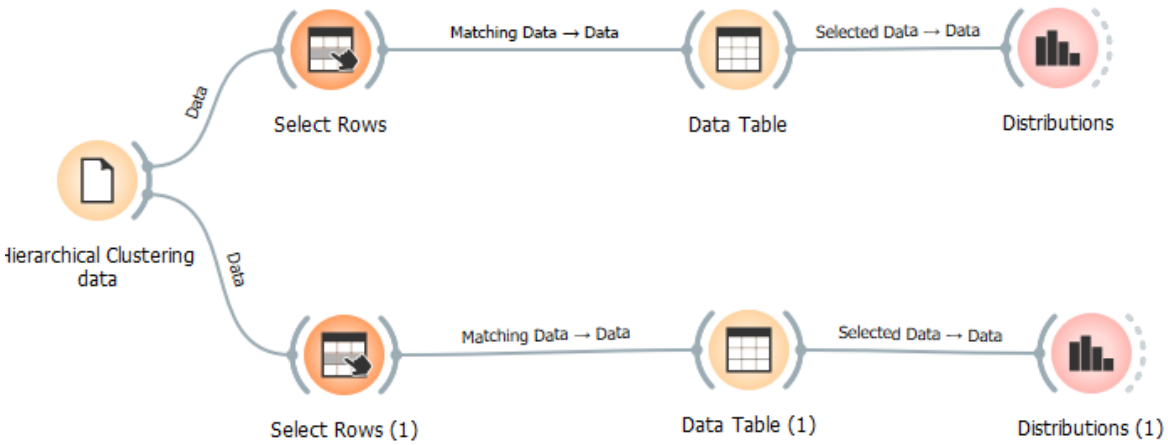
- Dựa vào chỉ số Silhouette Plot :

	Hierarchical Clustering	K-Means
C1	0.802	0.723
C2	0.666	0.688

Bảng 3. Bảng kết quả chỉ số Silhouette Plot

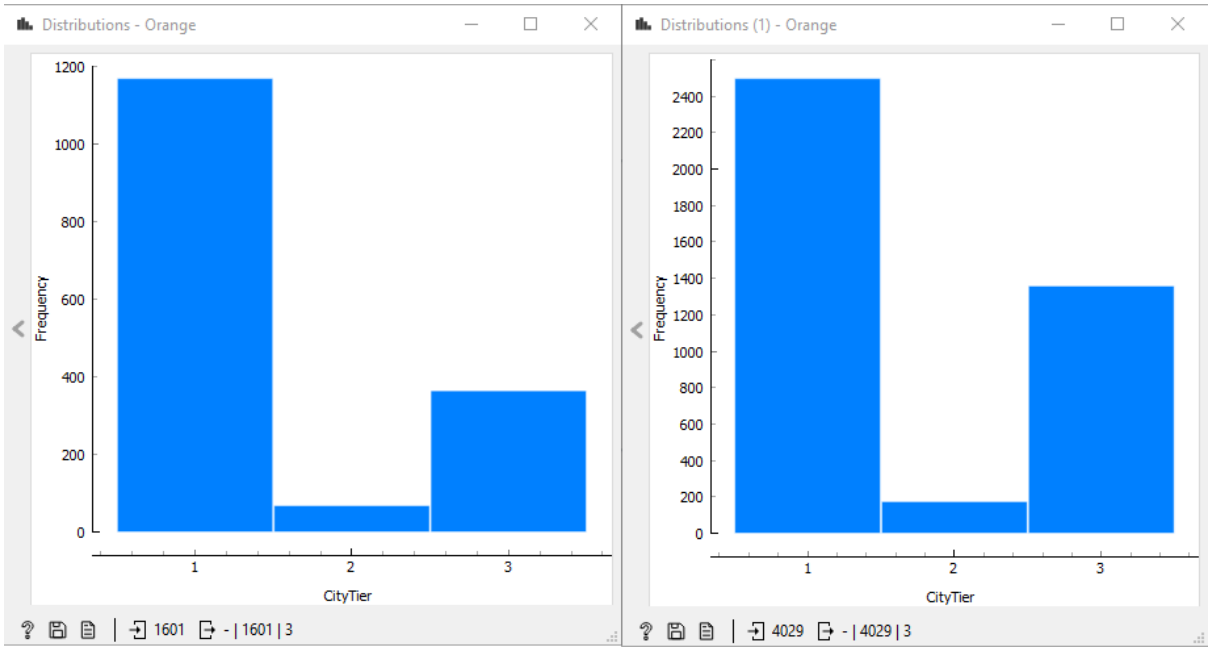
=> Nên chọn phương pháp Hierarchical Clustering vì có chỉ số Silhouette tiến về 1 hơn.

3.4.2. Phân tích đặc điểm từng cụm dữ liệu



Hình 39. Mô hình so sánh giữa 2 cụm

3.4.2.1. City Tier



Hình 40. Kết quả so sánh giữa 2 cụm về thuộc tính CityTier

Nhận xét: So sánh số lượng người ở mỗi cấp thành phố giữa cụm 1 với cụm 2 ta có:

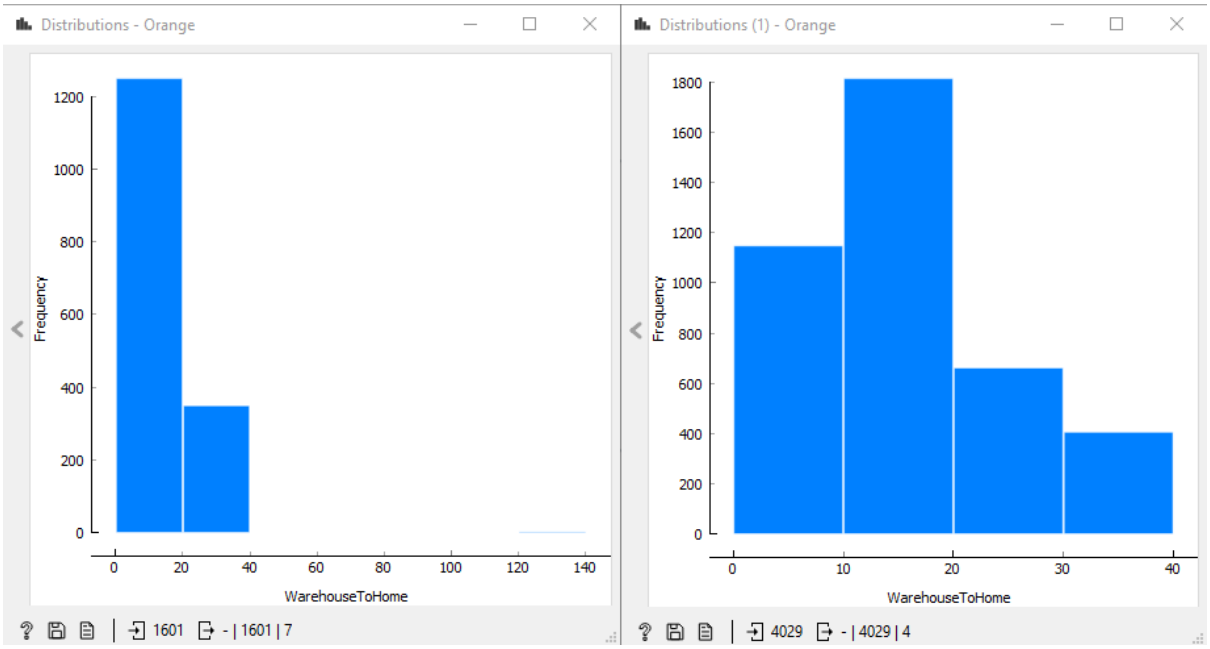
	Cấp 1	Cấp 2	Cấp 3
Cụm 1	1169	68	364
Cụm 2	2497	174	1358

Bảng 4. Bảng so sánh số lượng người ở mỗi cấp thành phố giữa cụm 1 với cụm 2

Từ đây thì ta có thể nhận thấy rằng:

- Ở cụm 1 số lượng người ở thành phố cấp 1 chiếm tỷ lệ cao nhất so với cấp 2 và cấp 3 (chiếm 73,02% trên tổng số). Thành phố cấp 2 có số lượng người ở là thấp nhất, chỉ chiếm 4,25%. Còn lại 22,74% là tỷ lệ người ở thành phố cấp 3.
- Còn ở cụm 2, số lượng người ở thành phố cấp 1 vẫn chiếm tỷ lệ cao nhất (61,98%), tiếp đến là số lượng người ở thành phố cấp 3 (33,71%), và thành phố cấp 2 có tỷ lệ người ở thấp nhất (4,32%).

3.4.2.2. Warehouse To Home



Hình 41. Kết quả so sánh giữa 2 cụm về thuộc tính WarehouseToHome

Nhận xét: So sánh về khoảng cách từ nhà kho đến nhà khách hàng của 2 cụm, ta có:

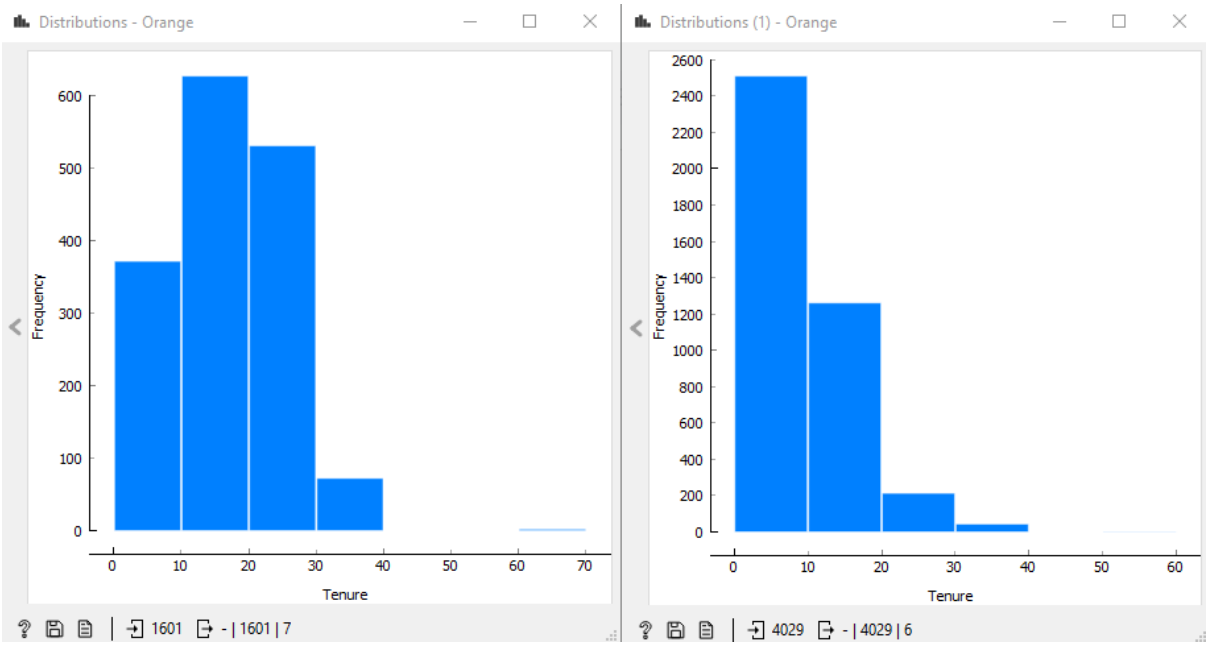
	Từ 0 đến 20 km	Từ 20 đến 40 km
Cụm 1	1250	349
Cụm 2	2963	1066

Bảng 5. Bảng so sánh về khoảng cách từ nhà kho đến nhà khách hàng của 2 cụm

Ta nhận thấy rằng:

- Ở cả 2 cụm, khoảng cách từ nhà kho đến nhà khách hàng chủ yếu phần lớn nằm trong khoảng từ 0 đến 20 km,
- Cụ thể hơn:
  + Ở cụm 1, số lượng là 1250 chiếm tỷ lệ 78,08% so với số còn lại là 21,08%.
  + Còn ở cụm 2, số lượng khách hàng nằm trong khoảng cách 0 đến 20 km là 2963 người, chiếm tổng cộng là 73,54% trong khi số còn lại chỉ chiếm tổng cộng 24,46%.

3.4.2.3. Tenure



Hình 42. Kết quả so sánh giữa 2 cụm về thuộc tính Tenure

Nhận xét: So sánh về thời gian khách hàng gắn bó với tổ chức giữa 2 cụm:

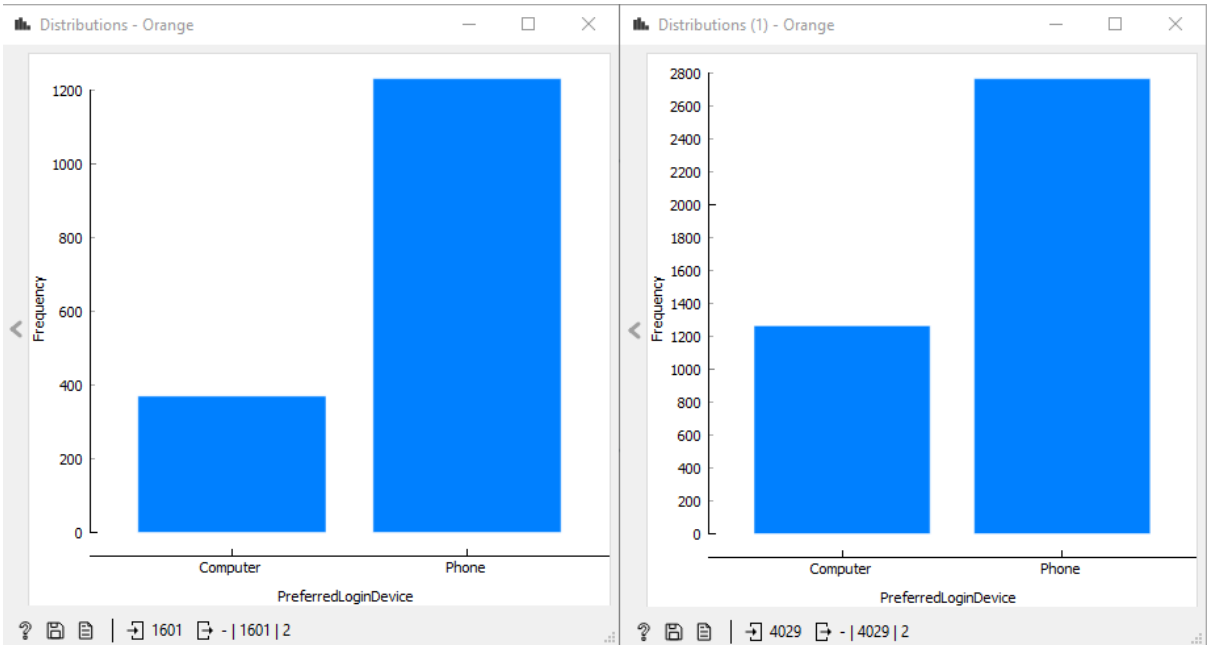
	0 - 10	10 - 20	20 - 30	30 - 40	Trên 40
Cụm 1	371	626	530	72	2
Cụm 2	2510	1261	213	43	2

Bảng 6. Bảng so sánh về thời gian khách hàng gắn bó với tổ chức giữa 2 cụm

Từ bảng trên, ta có thể thấy:

- Cụm số 1, phần lớn khách hàng có thời gian gắn bó với tổ chức nằm trong khoảng từ 10 đến 20 tháng (39,10%), tiếp đến là gắn bó từ 20 đến 30 tháng với 530 khách hàng (33,10%), sau đó sẽ là 371 người gắn bó từ 0 đến 10 tháng (23,17%), từ khoản 30 đến 40 tháng chiếm tỷ lệ khoảng (4,50%) và cuối cùng là có 2 khách hàng gắn bó với tổ chức trên 40 tháng (0,12%).
- Cụm số 2, phần lớn khách hàng có thời gian gắn bó với tổ chức sẽ nằm trong khoảng từ 0 đến 10 tháng (62,30%), tiếp đến là số lượng khách hàng gắn bó từ 10 đến 20 tháng (31,30%), sau đó là 213 người (5,29%) trong khoảng từ 20 đến 30 tháng, từ 30 đến 40 tháng có 43 khách hàng (1,07%), và sau cùng có 2 khách hàng đã gắn bó trên 40 tháng chiếm tỷ lệ 0,05%
- Giữa 2 cụm thì chủ yếu khách hàng của cụm số 2 là những khách hàng mới gắn bó với tổ chức trong khi cụm số 1 thì đã là những khách hàng tham gia lâu năm và trung thành với tổ chức.

3.4.2.4. PreferredLoginDevice



Hình 43. Kết quả so sánh giữa 2 cụm về thuộc tính PreferredLoginDevice

Nhận xét: So sánh về thiết bị đăng nhập ưa thích của khách hàng giữa 2 cụm:

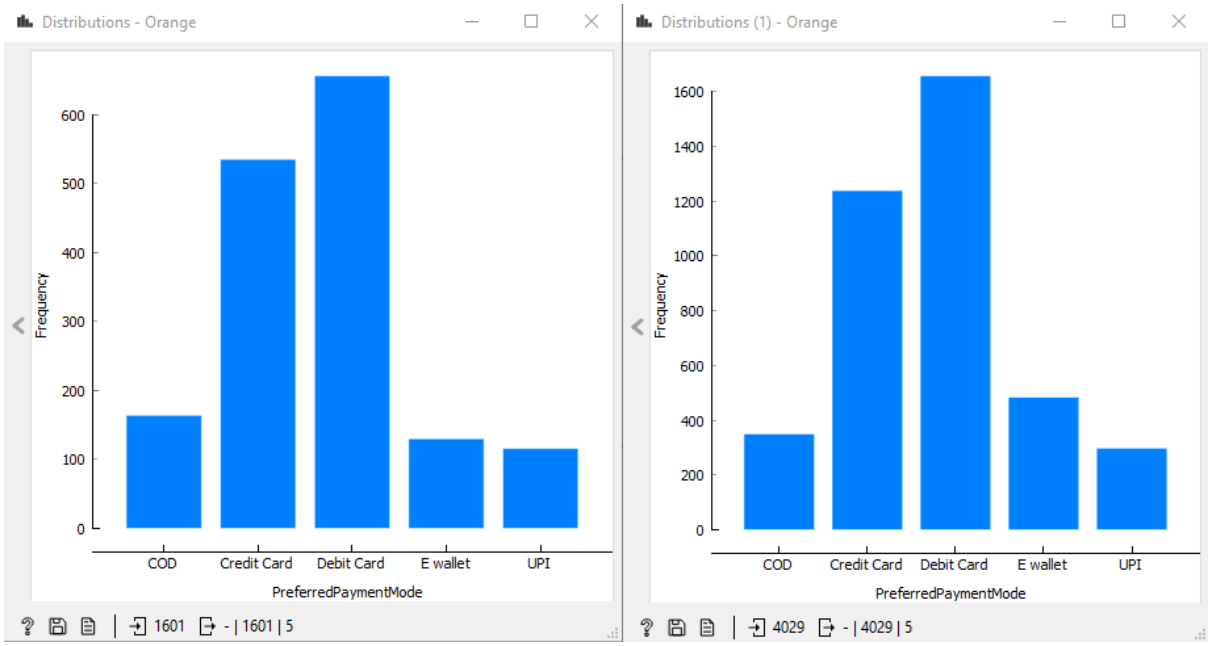
	Computer	Phone
Cụm 1	370	1231
Cụm 2	1246	2765

Bảng 7. Bảng so sánh về thiết bị đăng nhập ưa thích của khách hàng giữa 2 cụm

Kết quả trên cho thấy:

- Ở cụm 1: Điện thoại là thiết bị được khách hàng ưa thích lựa chọn để đăng nhập với 1231 người sử dụng, chiếm tỷ lệ 76,89%, gần gấp 4 lần so với sử dụng máy tính để đăng nhập (23,11%).
- Ở cụm 2: Giống với cụm 1, điện thoại là thiết bị được sử dụng nhiều nhất với 2765 khách hàng lựa chọn (68,63%), cao hơn việc sử dụng máy tính để đăng nhập (31,37%)
- Nhìn chung, khách hàng có thể ưa thích tính tiện dụng của điện thoại khi đăng nhập hơn nên việc sử dụng điện thoại đều chiếm ưu thế ở cả 2 cụm.

3.4.2.5. PreferredPaymentMode



Hình 44. Kết quả so sánh giữa 2 cụm về thuộc tính PreferredPaymentMode

Nhận xét: So sánh về hình thức thanh toán ưa thích của khách hàng giữa 2 cụm:

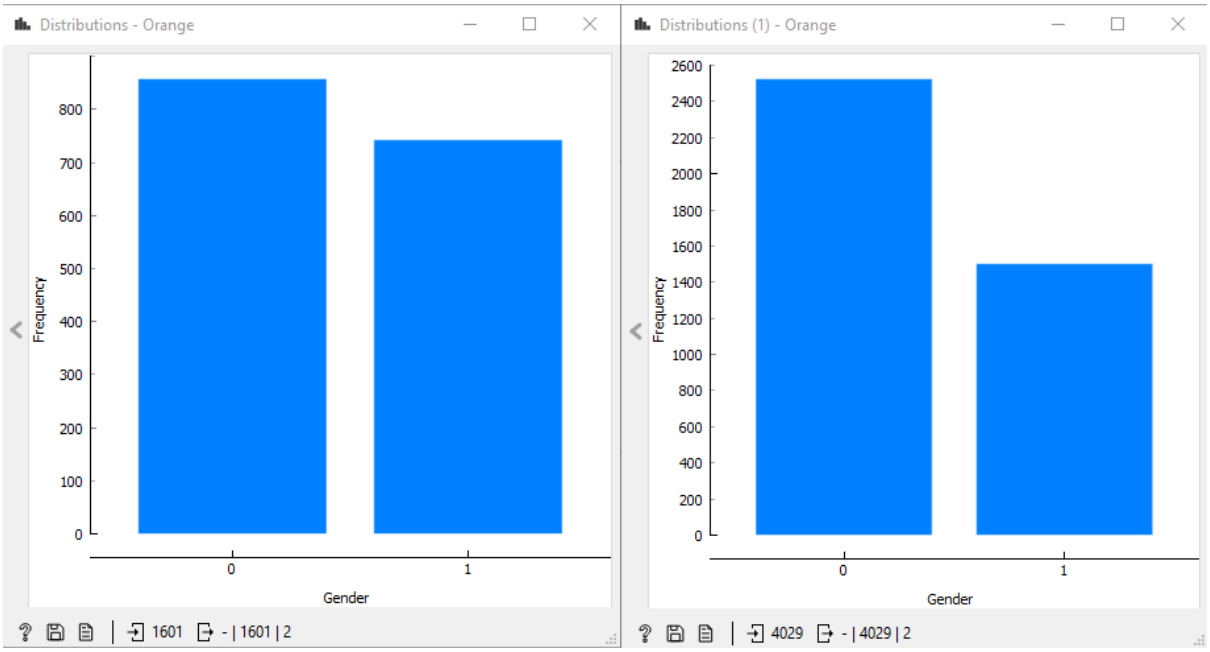
	COD	Credit Card	Debit Card	E-wallet	UPI
Cụm 1	164	535	656	130	116
Cụm 2	350	1239	1658	484	294

Bảng 8. Bảng so sánh hình thức thanh toán ưa thích của khách hàng giữa 2 cụm

Ta rút ra kết luận như sau:

- Cụm 1: Debit Card (thẻ ghi nợ) là phương thức được nhiều khách hàng ưa chuộng (40,97%). Tiếp đến là phương thức trả bằng Credit Card (thẻ tín dụng) với 33,42%. Các phương thức còn lại lần lượt là COD (10,24%), E-wallet (8,12%), UPI (7,25%).
- Cụm 2: Thanh toán bằng Debit Card (thẻ ghi nợ) vẫn là phương thức được nhiều khách hàng ưa thích với tỷ lệ 41,15%, tiếp đến vẫn là Credit Card với 30,75%. Các phương thức còn lại như COD, E-wallet, UPI lần lượt chiếm tỷ lệ lần lượt là: 8,69%; 12,01%; 7.40%. Trong đó có thể nhận thấy thì ở cụm 2, tỷ lệ dùng E-wallet cao hơn hẳn so với cụm 1.
- Các khách hàng ưa chuộng Debit Card và Credit Card thì 2 phương thức này có thể đáp ứng được sự tiện lợi khi mua hàng được phép chuyển tiền thanh toán trước, đồng thời 2 hình thức thanh toán này khá quen thuộc với khách hàng. Do đó, đây là những phương thức thanh toán có thể nói là tối ưu với khách hàng, so với việc trả khi nhận hàng (COD) hoặc những phương thức mới hiện đại hơn như ví điện tử (E-wallet), UPI,..

3.4.2.6. Gender



Hình 45. Kết quả so sánh giữa 2 cụm về thuộc tính Gender

Nhận xét: So sánh về giới tính của khách hàng giữa 2 cụm:

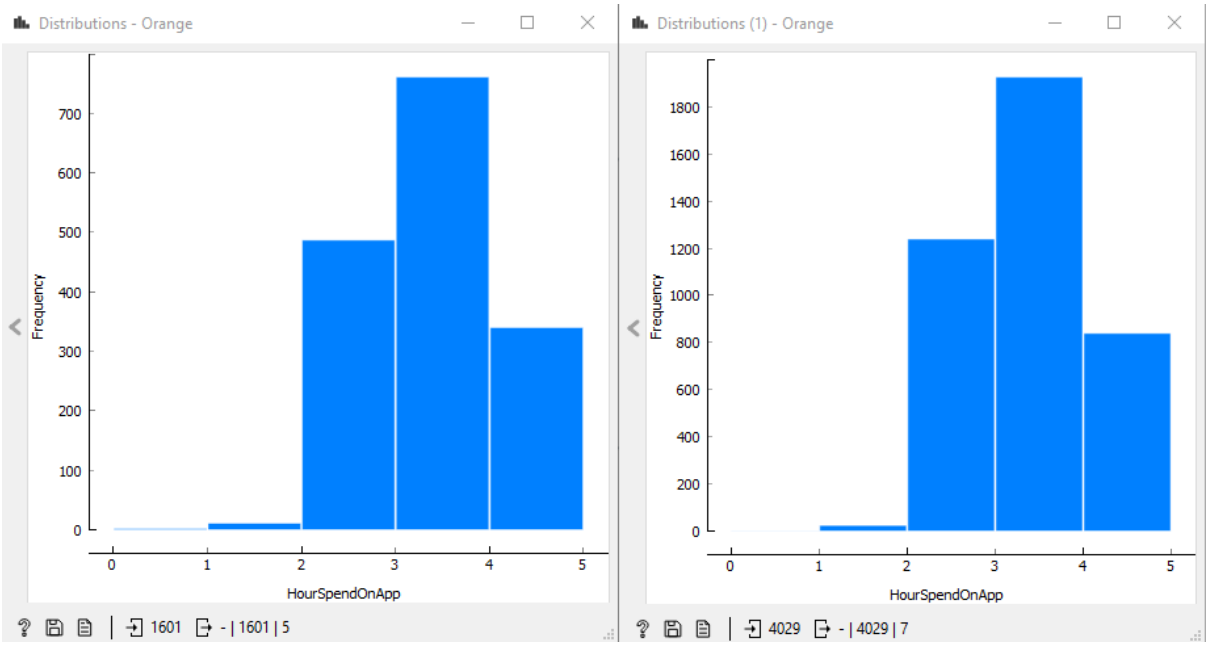
	Nam (0)	Nữ (1)
Cụm 1	858	743
Cụm 2	2526	1503

Bảng 9. Bảng so sánh giới tính của khách hàng giữa 2 cụm

Ta rút ra kết luận như sau:

- Cụm 1: Số lượng khách hàng giới tính nam (53,59%) chiếm số lượng nhiều hơn so với số lượng khách hàng giới tính là nữ (46,41%).
- Cụm 2: Giống như cụm 1, khách hàng giới tính nam (62,70%) vẫn chiếm ưu thế hơn khách hàng giới tính nữ (37,30%).
- Nhìn chung, theo như dữ liệu, ta nhận thấy rằng khách hàng có giới tính nam chiếm phần nhiều hơn.

3.4.2.7. HourSpendOnApp



Hình 46. Kết quả so sánh giữa 2 cụm về thuộc tính HourSpendOnApp

Nhận xét: So sánh về thời gian khách hàng dành ra để lướt app hoặc web khách hàng giữa 2 cụm:

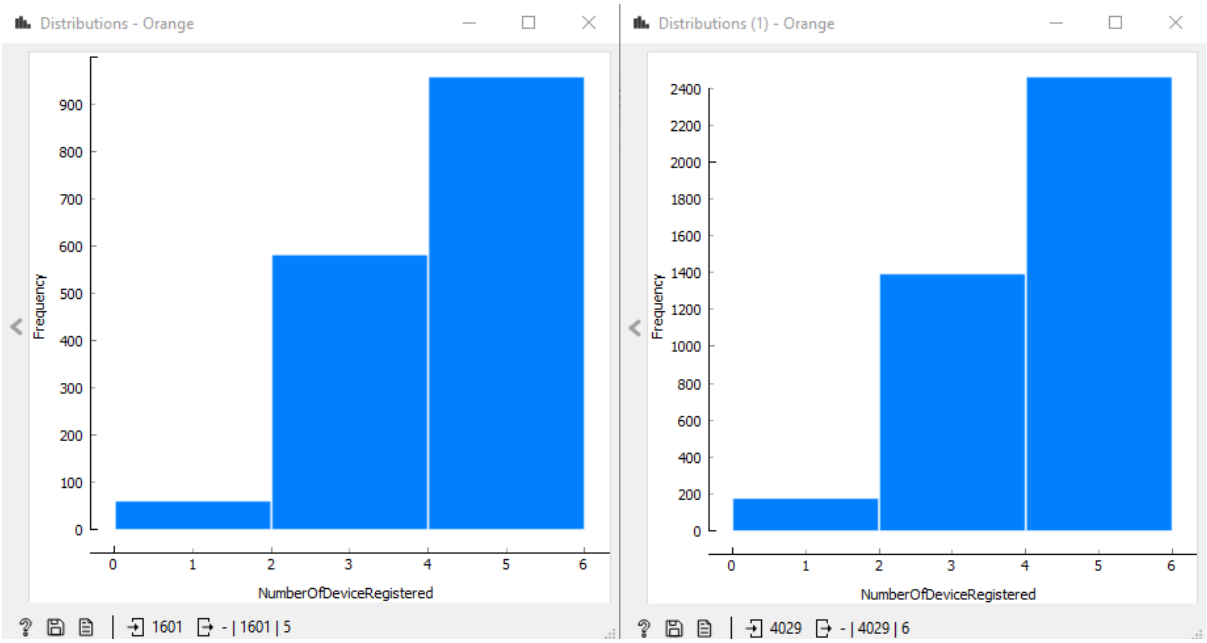
	Dưới 1h	1-2h	2-3h	3-4h	Hơn 4h
Cụm 1	2	11	487	761	340
Cụm 2	1	24	1239	1926	839

Bảng 10. Bảng so sánh thời gian khách hàng dành ra để lướt app hoặc web khách hàng giữa 2 cụm

Từ bảng kết quả trên, ta rút ra:

- Ở cụm 1: Chủ yếu những khách hàng ở cụm 1 sẽ dành ra khoảng từ 3 đến 4 giờ đồng hồ mỗi ngày để lướt app hoặc web (47,53%), tiếp đó là khoảng từ 2 đến 3h mỗi ngày (30,42%) và hơn 4 giờ mỗi ngày (21,24%). Chỉ có 11 khách hàng dành 1 đến 2 giờ để lướt app hoặc web và 2 khách hàng chỉ dành dưới 1 giờ mỗi ngày.
- Ở cụm 2: Có 1926 khách hàng đã dành 3 đến 4 giờ mỗi ngày để lướt app hoặc web (47,80%), 1239 khách hàng dành ra 2 đến 3 giờ (30,75%) và 839 khách hàng dành hơn 4h cho việc lướt app/web (20,82%). Số ít khách hàng còn lại dành ra khoảng từ 1 đến 2h mỗi ngày hoặc dưới 1 giờ cho việc này.
- Tổng quan, đại đa số khách hàng họ dành trung bình khoảng từ 2 đến 4 giờ mỗi ngày cho việc lướt app hoặc web.

3.4.2.8. NumberOfDeviceRegistered



Hình 47. Kết quả so sánh giữa 2 cụm về thuộc tính NumberOfDeviceRegistered

Nhận xét: So sánh về tổng số thiết bị mà một khách hàng đăng ký giữa 2 cụm:

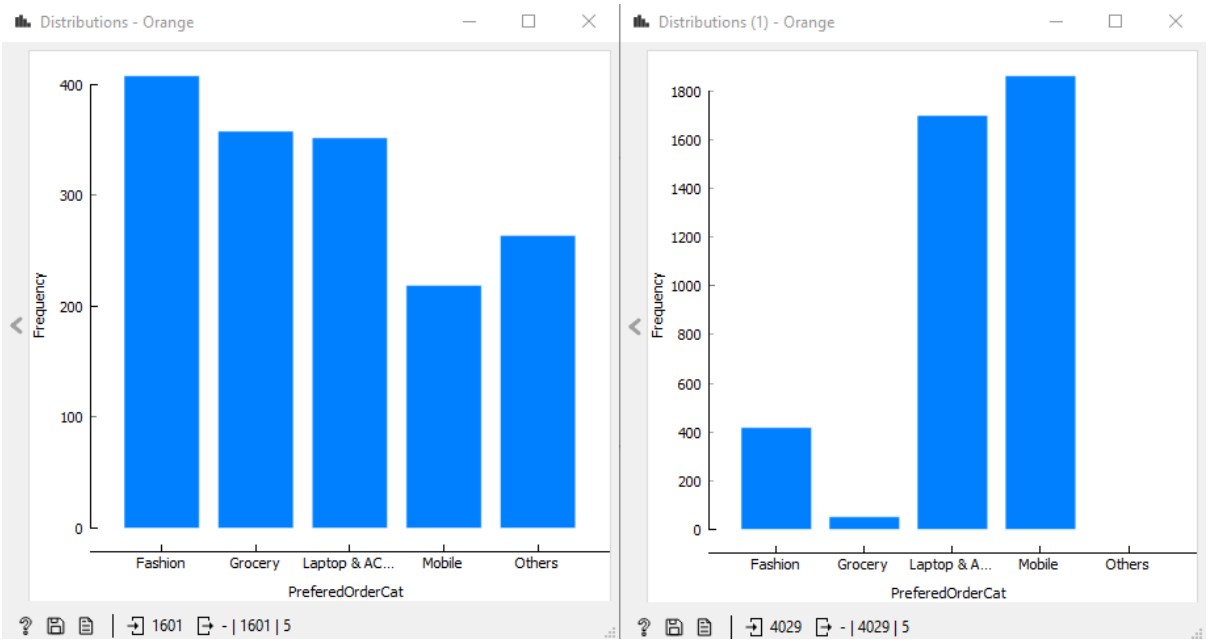
	dưới 2 thiết bị	2-4 thiết bị	nhiều hơn 4 thiết bị
Cụm 1	60	582	959
Cụm 2	175	1393	2416

Bảng 11. Bảng so sánh tổng số thiết bị mà một khách hàng đăng ký giữa 2 cụm

Ta có nhận xét như sau:

- Cụm 1 và cụm 2 giống nhau là chủ yếu các khách hàng dùng nhiều hơn 4 thiết bị để đăng ký. Ở cụm 1, có 959 cá nhân (59,90%) và cụm 2 có 2416 cá nhân (61,08%) dùng trên 4 thiết bị . Ngoài ra thì số lượng khách hàng có dùng dưới 2 thiết bị đăng ký là thấp nhất với 3,75% ở cụm 1 và 4,34% ở cụm 2. Còn lại là trong khoảng từ 2 đến 4 thiết bị.
- Có thể khách hàng dùng nhiều thiết bị với mong muốn có thêm những ưu đãi/khuyến mãi từ sàn thương mại điện tử, hoặc để dễ dàng xem xét và so sánh giá cả giữa những người bán hàng. Ngoài ra, còn có thể là do khách hàng cảm thấy tiện lợi và dễ dàng truy cập sàn thương mại điện tử hơn từ nhiều thiết bị khác nhau.

3.4.2.9. PreferedOrderCat



Hình 48. Kết quả so sánh giữa 2 cụm về thuộc tính PreferedOrderCat

Nhận xét: So sánh về Danh mục sản phẩm mà khách hàng ưa thích đặt tháng trước giữa 2 cụm:

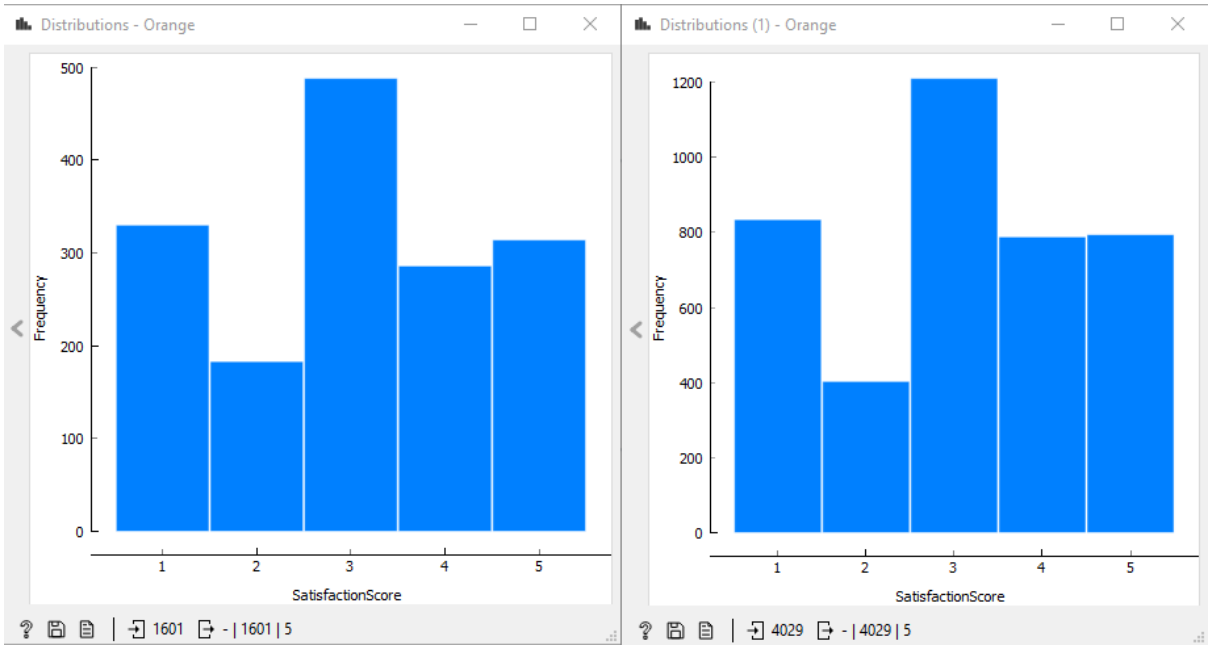
	Fashion	Grocery	Laptop & Accessory	Mobile	Others
Cụm 1	408	358	352	219	264
Cụm 2	418	52	1698	1861	0

Bảng 12. Bảng so sánh Danh mục sản phẩm mà khách hàng ưa thích đặt tháng trước giữa 2 cụm

Ta có thể thấy rằng:

- Cụm 1 không có quá nhiều sự chênh lệch giữa các danh mục. Trong đó, thời trang là danh mục sản phẩm mà khách hàng ưa thích đặt tháng trước nhất với 408 khách hàng lựa chọn (25,48%), tiếp đến là mặt hàng tạp hoá (22,36%) và mặt hàng laptop (21,99%). Hàng hoá khác và di động là 2 danh mục chiếm tỷ lệ thấp nhất, lần lượt là 16,49% và 13,68%.
- Cụm 2 cho thấy rằng khách hàng có xu hướng mua hàng công nghệ, bởi vì số liệu cao nhất của cụm 2 nằm ở danh mục Di động (46,19%) và danh mục Laptop (42,14%). Tiếp đến là Thời trang với tỷ lệ 10,37%. Hàng tạp hoá có tỷ lệ thấp nhất 1,29% và ngoài ra cụm 2 không có khách hàng nào chọn hàng hoá khác.

3.4.2.10. SatisfactionScore



Hình 49. Kết quả so sánh giữa 2 cụm về thuộc tính SatisfactionScore

Nhận xét: So sánh về điểm số hài lòng của khách hàng giữa 2 cụm:

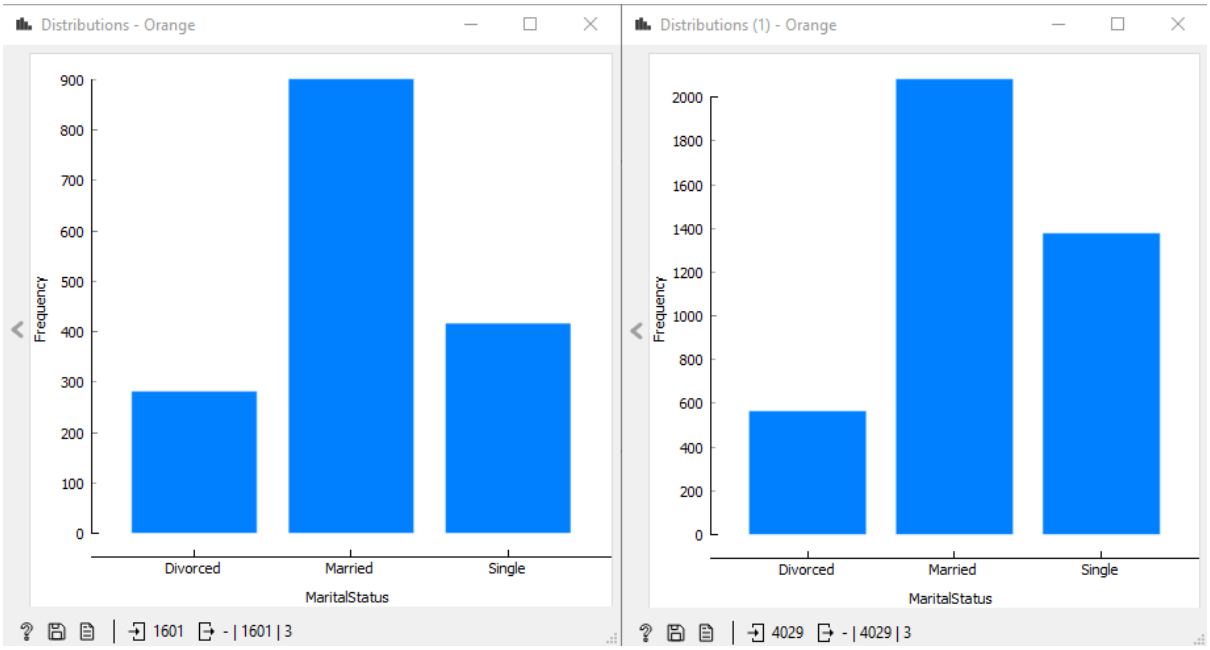
	1	2	3	4	5
Cụm 1	330	183	488	286	314
Cụm 2	834	403	1210	788	794

Bảng 13. Bảng so sánh điểm số hài lòng của khách hàng giữa 2 cụm

Ta thấy rằng:

- Ở cả 2 cụm, phần lớn khách hàng cảm thấy trung lập với điểm đánh giá là 3 điểm, trong đó cụm 1 có 488 khách hàng đánh giá 3 điểm (30,48%), cụm 2 có 1210 khách hàng (30,03%).
- Điểm 1 đều đứng thứ hai ở 2 cụm, lần lượt ở cụm 1 có 330 khách hàng và cụm 2 là 834 khách hàng đánh giá 1 điểm.
- Điểm 5 đứng thứ 3 với 19,61% ở cụm 1 và 19,71% ở cụm 2. Tiếp đến là 4 điểm với 17,86% ở cụm 1 và 19,56% ở cụm 2. Cuối cùng, ở cả 2 cụm, 2 điểm là điểm số có ít khách hàng đánh giá nhất.
- Nhìn chung, khách hàng chưa thực sự hài lòng với những dịch vụ mà sàn thương mại điện tử đang mang lại cho họ. Vì thế, doanh nghiệp nên có phương hướng để phát triển về dịch vụ khách hàng nhiều hơn trong tương lai.

3.4.2.11. MaritalStatus



Hình 50. Kết quả so sánh giữa 2 cụm về thuộc tính MaritalStatus

Nhận xét: So sánh về tình trạng hôn nhân của khách hàng giữa 2 cụm:

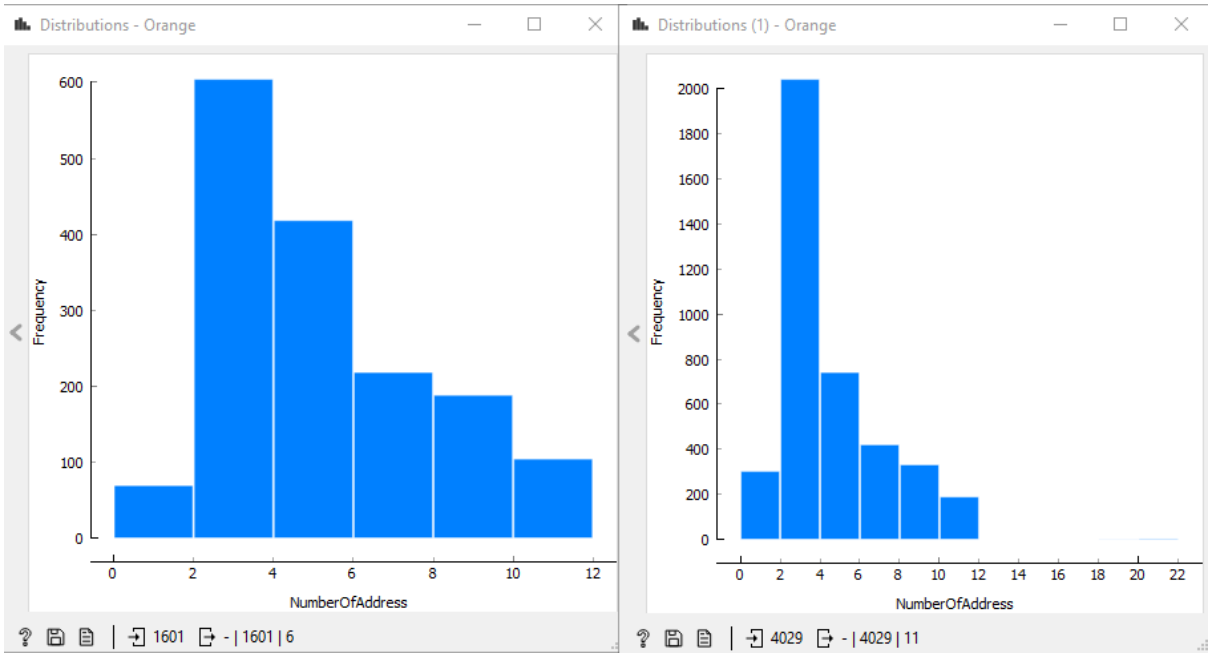
	Ly hôn	Kết hôn	Độc thân
Cụm 1	282	902	417
Cụm 2	566	2048	1379

Bảng 13. Bảng so sánh tình trạng hôn nhân của khách hàng giữa 2 cụm

Ta có thể thấy từ bảng trên:

- Ở cả cụm 1 và cụm 2, khách hàng phần lớn là người đã kết hôn, trong đó cụm 1 có 902 khách hàng (56,34%) và cụm 2 có 2048 khách hàng (51,72%)
- Tiếp đến, là những khách hàng còn độc thân với 417 khách hàng ở cụm 1 (26.05%) và 1379 khách hàng ở cụm 2 (34,23%).
- Cuối cùng là những khách hàng đã ly hôn chiếm tỷ lệ thấp nhất với 14,05% ở cụm 1 và 17,61% ở cụm 2.

3.4.2.12. NumberOfAddress



Hình 51. Kết quả so sánh giữa 2 cụm về thuộc tính NumberOfAddress

Nhận xét: So sánh về tổng số lượng địa chỉ mà một khách hàng đăng ký giữa 2 cụm:

	0-2	2-4	4-6	6-8	8-10	Trên 10
Cụm 1	69	604	418	218	188	104
Cụm 2	302	2043	741	420	331	191

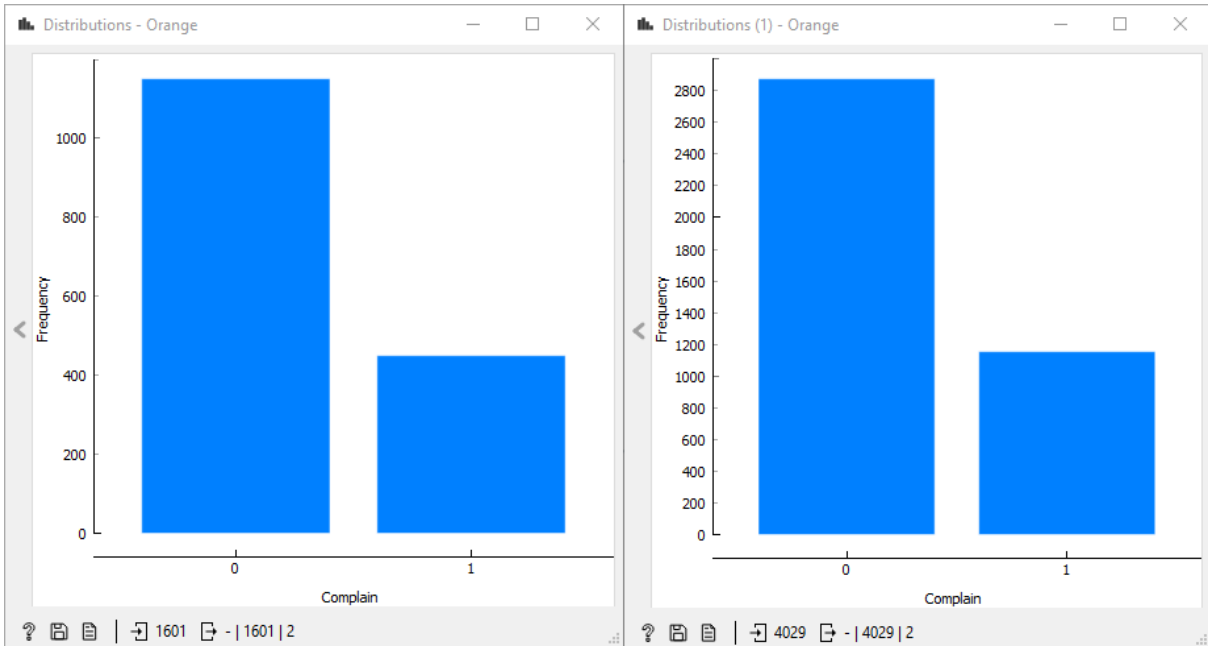
Bảng 14. Bảng so sánh tổng số lượng địa chỉ mà một khách hàng đăng ký giữa 2 cụm

Ta có thể rút ra nhận xét:

- Ở cụm 1, phần lớn khách hàng có số lượng địa chỉ đăng ký nằm trong khoảng 2-4 (37,73%). Tiếp đến là trong khoảng từ 4-6 địa chỉ (13,62%). 6-8 có tỷ lệ khoảng 13,26%, 8-10 địa chỉ có tỷ lệ 11,74% và trên 10 địa chỉ có 104 khách hàng với tỷ lệ 6,5%. Khách hàng có 0-2 địa chỉ đăng ký chiếm tỷ lệ nhỏ nhất với 4,31%.
- Ở cụm 2, phần lớn khách hàng cũng có khoảng 2-4 địa chỉ đăng ký, chiếm tỷ lệ cao nhất với 50,71%. Kế đến là trong khoảng từ 4-6 địa chỉ chiếm khoảng 18,39%. 6-8 địa chỉ chiếm tỷ lệ 10,42%, 8-10 địa chỉ chiếm 8,22% và 0-2 địa chỉ

có tỷ lệ 7,50%. Tỷ lệ trên 10 địa chỉ đăng ký là thấp nhất cụm với 191 khách hàng, tỷ lệ 4,74%.

3.4.2.13. Complain



Hình 52. Kết quả so sánh giữa 2 cụm về thuộc tính Complain

Nhận xét: So sánh về lời phàn nàn từ khách hàng trong tháng trước giữa 2 cụm:

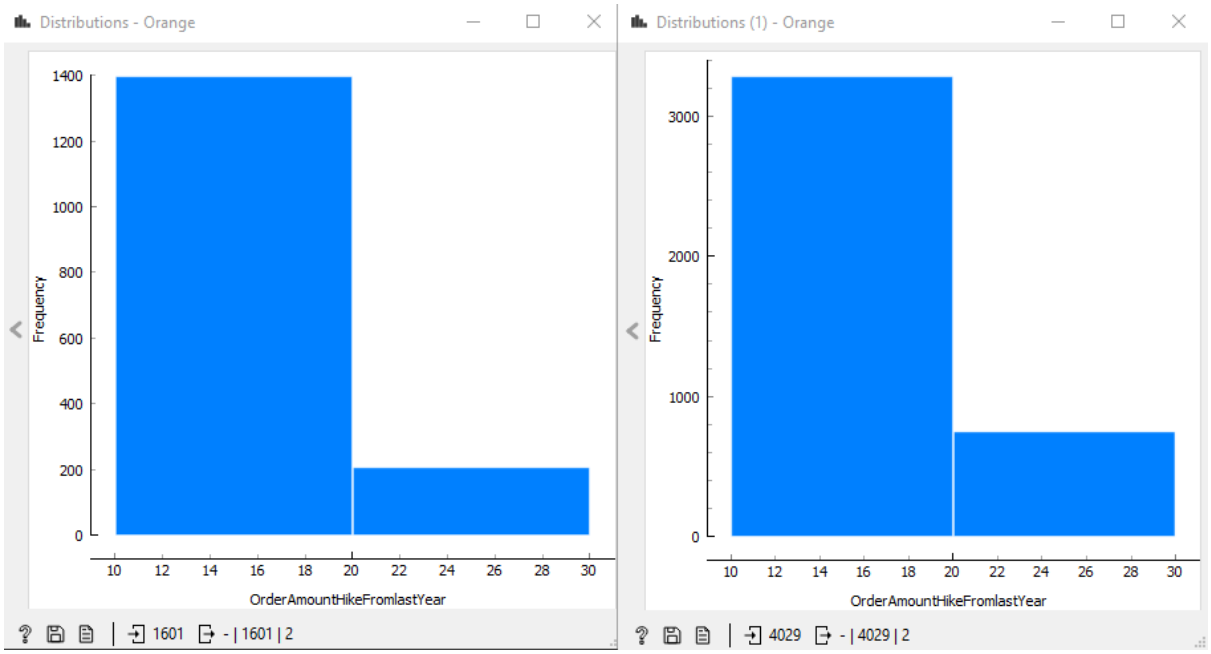
	Không (0)	Có (1)
Cụm 1	1151	450
Cụm 2	2875	1154

Bảng 15. Bảng so sánh lời phàn nàn từ khách hàng trong tháng trước giữa 2 cụm

Từ bảng số liệu trên, ta nhận thấy rằng:

- Cụm 1 có 1151 khách hàng không có lời phàn nàn nào vào tháng trước (71,89%) và có 450 khách hàng phàn nàn vào tháng vừa qua (28,11%).
- Cụm 2 có 2875 khách hàng tháng trước không phàn nàn (71,36%) và có 1154 khách hàng đã phàn nàn (28,64%).
- Nhìn chung, tuy rằng tỷ lệ không nhận được phàn nàn vẫn cao hơn nhưng cũng không có ít những sự không hài lòng và phàn nàn của khách hàng. Doanh nghiệp nên có những kế hoạch về cải tiến về công nghệ và dịch vụ chăm sóc khách hàng tốt hơn để hạn chế việc khiến khách hàng không vừa ý và phải đưa ra lời phàn nàn trong những tháng tiếp theo.

3.4.2.14. OrderAmountHikeFromlastYear



Hình 53. Kết quả so sánh giữa 2 cụm về thuộc tính OrderAmountHikeFromlastYear

Nhận xét: So sánh về phần trăm tăng trưởng đặt hàng trong năm trước giữa 2 cụm:

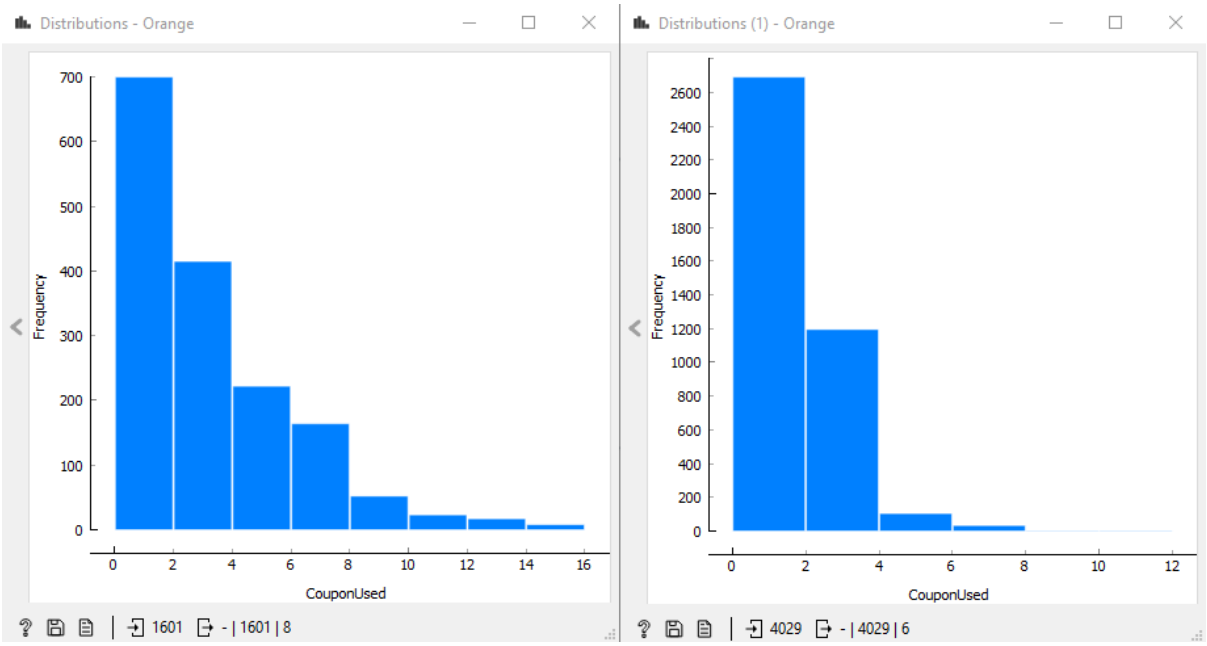
	Dưới 20%	Trên 20%
Cụm 1	1396	205
Cụm 2	3283	746

Bảng 16. Bảng so sánh phần trăm tăng trưởng đặt hàng trong năm trước giữa 2 cụm

Từ dữ liệu trên:

- Cụm 1 có 1396 khách hàng có phần trăm tăng trưởng đặt hàng năm ngoái dưới 20% (87,20%), chiếm ưu thế hơn so với 205 khách hàng có phần trăm tăng trưởng đặt hàng năm ngoái trên 20% (12,80%)
- Cụm 2 có 3283 khách hàng có phần trăm tăng trưởng đặt hàng vào năm ngoái dưới 20%, chiếm tỷ lệ 81,48% so với tỷ lệ 18,52% của nhóm khách hàng có phần trăm tăng trưởng đặt hàng năm trước trên 20%.
- Tổng quan thì đây vẫn là một tỷ lệ ổn định. Tuy nhiên, về lâu dài trong tương lai, lời khuyên cho sàn thương mại điện tử này là có những chiến lược tiếp thị thu hút hơn để duy trì và gia tăng số lượng phần trăm tăng trưởng đặt hàng của khách hàng.

3.4.2.15. CouponUsed



Hình 54. Kết quả so sánh giữa 2 cụm về thuộc tính CouponUsed

Nhận xét: So sánh về tổng số coupon đã sử dụng trong tháng trước giữa 2 cụm:

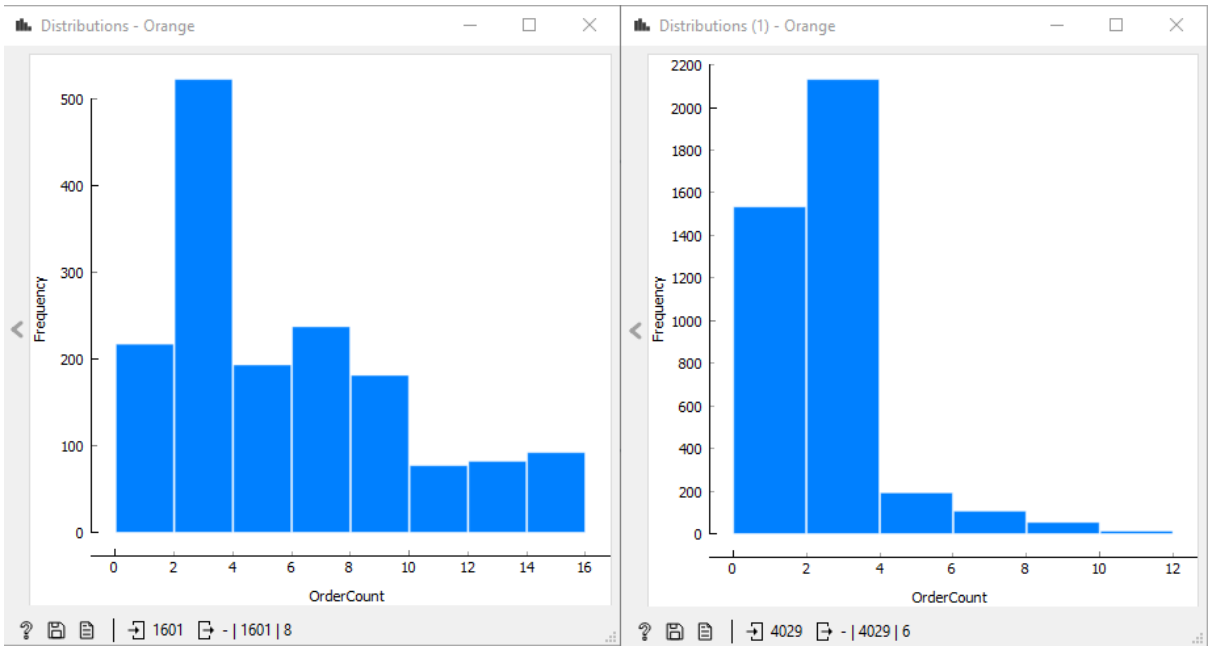
	0-2	2-4	4-6	6-8	8-10	Trên 10
Cụm 1	700	415	222	164	52	48
Cụm 2	2691	1195	104	33	3	3

Bảng 17. Bảng so sánh tổng số coupon đã sử dụng trong tháng trước giữa 2 cụm

Ta nhận thấy rằng:

- Ở cụm 1, khách hàng phần lớn sử dụng từ 0 đến 2 coupon vào tháng vừa rồi, tỷ lệ 43,72%. Đứng thứ 2 là trong khoảng 2 đến 4 coupon (25,92%). Kế đến là trong khoảng 4-6 coupon với 13,87% và 6-8 coupon với 10,24%. Trong khoảng 8-10 coupon và trên 10 coupon có dưới 100 khách hàng, cụ thể 8-10 có 52 khách hàng và thấp nhất là sử dụng trên 10 với 48 khách hàng.
- Ở cụm 2, ta dễ nhận thấy có 1 sự chênh lệch khá lớn giữa các khoảng. Khách hàng phần lớn cũng sử dụng trong khoảng 0-2 coupon trong tháng trước, chiếm tỷ lệ lớn nhất là 66,79%, tiếp đến là trong khoảng từ 2-4 với 1195 khách hàng, chiếm tỷ lệ (29,66%). Từ 4-6 có 104 khách hàng, tỷ lệ 2,58%. Không có quá nhiều khách hàng ở cụm 2 tháng vừa rồi sử dụng coupon trong các khoảng 6-8,8-10 và trên 10, tỷ lệ chiếm rất nhỏ, cụ thể lần lượt là: 0,82%, 0,07% và 0,07%.
- Tóm lại, đại đa số những khách hàng tham gia vào sàn thương mại điện tử chỉ sử dụng trung bình từ 0 đến 4 coupon trong 1 tháng. Từ dữ liệu về số lượng coupon được khách hàng thì ta cũng dự đoán được số lượng đơn đặt hàng trung bình của phần lớn khách hàng.

3.4.2.16. OrderCount



Hình 55. Kết quả so sánh giữa 2 cụm về thuộc tính OrderCount

Nhận xét: So sánh về tổng số đơn hàng được đặt trong tháng trước giữa 2 cụm:

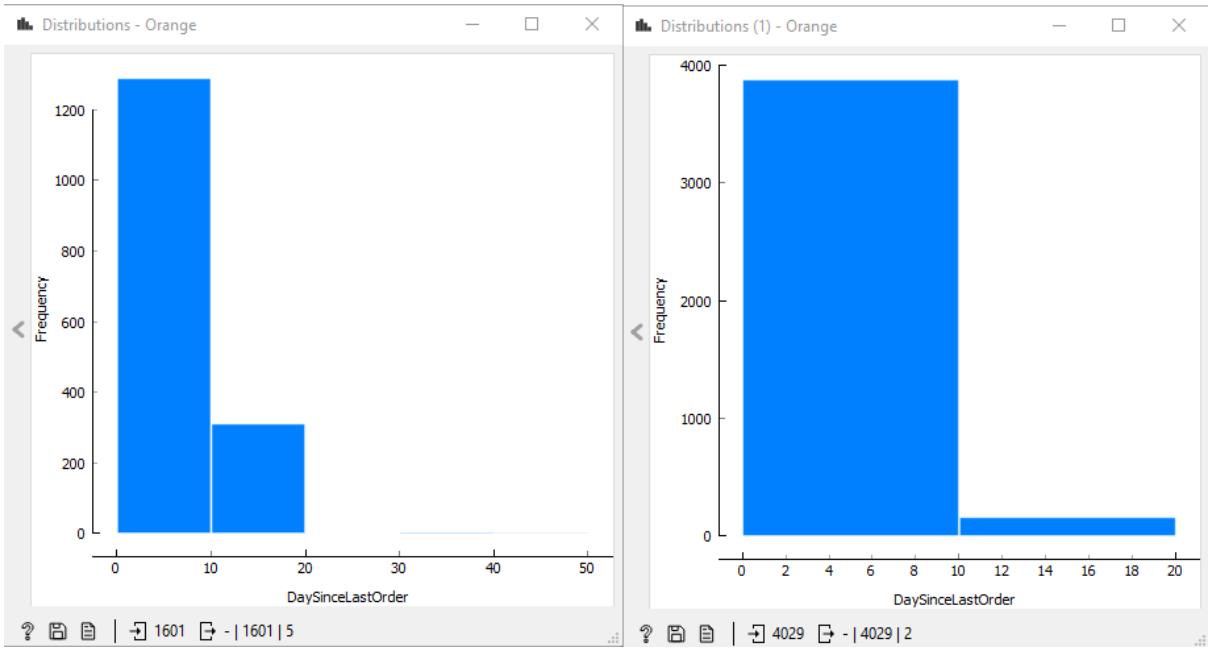
	0-2	2-4	4-6	6-8	8-10	Trên 10
Cụm 1	217	522	193	273	181	251
Cụm 2	1534	2132	192	106	53	12

Bảng 18. Bảng so sánh tổng số đơn hàng được đặt trong tháng trước giữa 2 cụm

Từ bảng trên:

- Cụm 1 có tỷ lệ trải đều hơn so với cụm thứ 2. Trong đó, chiếm ưu thế là trong khoảng từ 2-4 đơn hàng vào tháng trước (31,60%). Những khoảng còn lại không có quá nhiều những số liệu quá chênh lệch nhau.
- Cụm 2 có đại đa số khách hàng đặt khoảng từ 2-4 đơn hàng với tỷ lệ 52,92%, kế đến là trong khoảng từ 0-2 đơn hàng với 1534 khách hàng, tỷ lệ 38,07%. Các khoảng còn lại thì khá thấp.
- Từ đó, ta thấy rằng ở cụm số 1, khách hàng có hành vi mua hàng trên sàn thương mại điện tử thường xuyên hơn là cụm số 2.

3.4.2.17. DaySinceLastOrder



Hình 56. Kết quả so sánh giữa 2 cụm về thuộc tính DaySinceLastOrder

Nhận xét: So sánh về ngày mà lần cuối đặt hàng giữa 2 cụm:

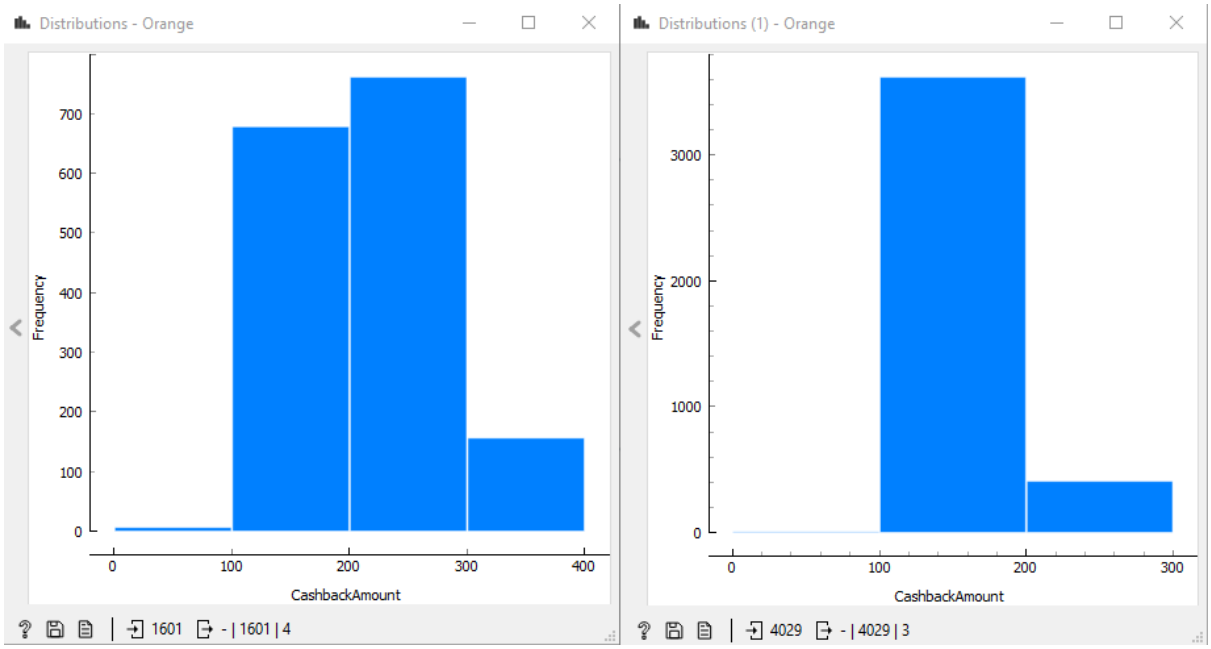
	Dưới 10 ngày	Trên 10 ngày
Cụm 1	1289	312
Cụm 2	3876	153

Bảng 19. Bảng so sánh ngày mà lần cuối đặt hàng giữa 2 cụm

Ta nhận thấy rằng:

- So sánh cả 2 cụm, ta thấy rằng ở cụm 2 có đại đa số khách hàng có dưới 10 ngày từ lần cuối đặt hàng (96,20%), so với cụm 1 (80,51%).
- Số khách hàng có trên 10 ngày từ ngày đặt hàng lần cuối ở cụm 1 nhiều hơn so với cụm 2.
- Vậy, cụm 2 sẽ có nhiều khách hàng thường xuyên truy cập sàn thương mại điện tử để mua hàng hơn là so với cụm 1.

3.4.2.18. CashbackAmount



Hình 57. Kết quả so sánh giữa 2 cụm về thuộc tính CashbackAmount

Nhận xét: So sánh về trung bình tiền trả lại tháng trước giữa 2 cụm (đvt: $)

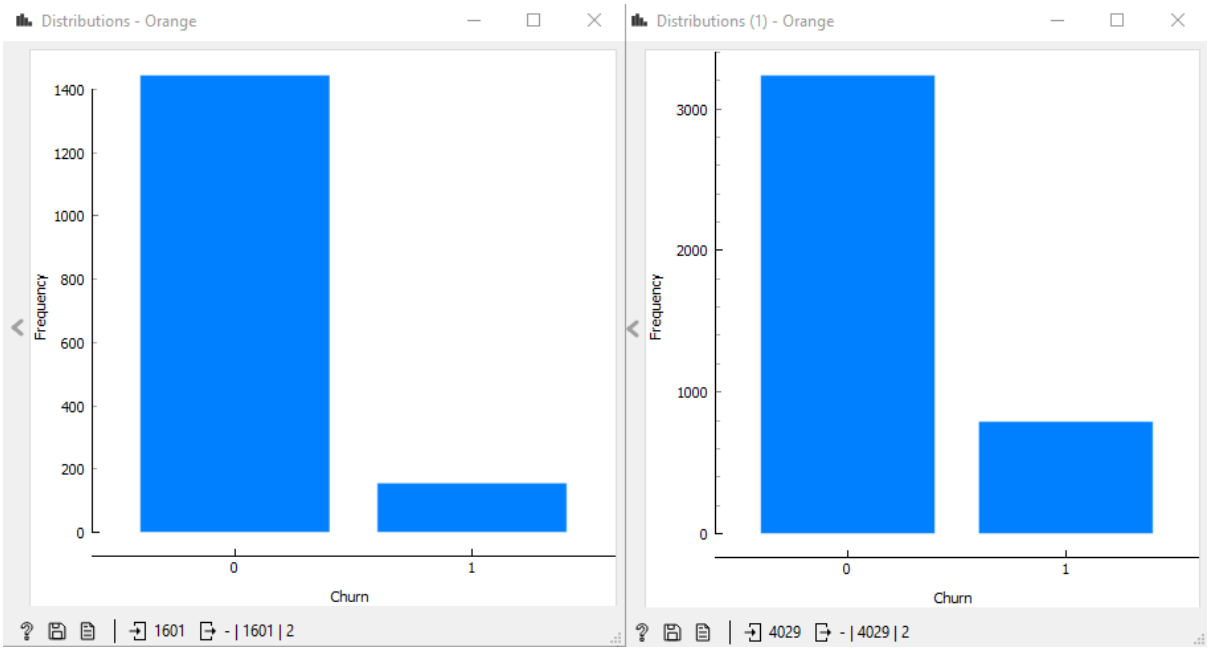
	0-100	100-200	200-300	Trên 300
Cụm 1	6	678	761	156
Cụm 2	6	3617	406	0

Bảng 20. Bảng so sánh trung bình tiền trả lại tháng trước giữa 2 cụm (đvt: $)

Từ bảng tổng hợp trên:

- Ta nhận thấy rằng cụm 1 có nhiều khách hàng có số tiền trả lại trung bình trên 300 trong khi cụm số 2 lại không có khách hàng nào.
- Trong khoảng 200$ đến 300$ là khoảng mà cụm 1 có tỷ lệ cao nhất (47,53%), tiếp đến là trong khoảng 100$ đến 200$ với 678 khách hàng (42,35%) và trên 300% có tỷ lệ 9,74%. Từ 0$ đến 100$ có tỷ lệ ít nhất.
- Cụm 2, ta có thể thấy rằng phần lớn khách hàng có số tiền trả lại tháng trước nằm trong khoảng từ 100$ đến 200$ (89,77%). Tiếp đến là trong khoảng từ 200$ đến 300$ (10,08%) và 6 khách hàng có số tiền từ 0$ đến 100$ (0,15%). Không có khách hàng nào có số tiền hoàn trả trên 300$.

3.4.2.19. Churn



Hình 58. Kết quả so sánh giữa 2 cụm về thuộc tính Churn

Nhận xét: So sánh về khách hàng rời bỏ dịch vụ giữa 2 cụm:

	Không (0)	Có (1)
Cụm 1	1445	156
Cụm 2	3257	792

Bảng 21. Bảng so sánh khách hàng rời bỏ dịch vụ giữa 2 cụm

Nhìn chung, ta thấy:

- Số lượng khách hàng lựa chọn Không rời bỏ dịch vụ ở 2 cụm đều chiếm tỷ lệ phần lớn, lần lượt là cụm 1 với 1445 khách hàng (90,25%) và cụm 2 với 3257 khách hàng (80,34%).
- Tuy nhiên thì những con số này sẽ có thể thay đổi, vì thế nên sàn thương mại điện tử nên có một kế hoạch hoặc chiến lược cụ thể và lâu dài để tiếp tục duy trì khách hàng cũ

3.5. Đề xuất phương hướng phát triển từ mô hình phân cụm

Sau khi phân tích các thuộc tính của bộ dữ liệu từ mô hình phân cụm, nhóm xin đề xuất một số phương hướng phát triển cho sàn thương mại điện tử trong tương lai, cụ thể như sau:

3.5.1. Nhận xét phân cụm

Từ kết quả phân 2 cụm đã đề cập ở trên, có thể thấy rằng các đặc điểm riêng của mỗi cụm như sau:

CỤM 1	CỤM 2
- Khách hàng lâu năm, trung thành với hệ thống. - Tỉ lệ Nam-Nữ gần bằng nhau (khoảng 50-50) nên những nhu cầu mua sắm sẽ đa dạng hơn, cụ thể cao nhất là Thời Trang. - Chất lượng nơi ở khá cao (được thể hiện thông qua CityTier) khi phần nhiều là ở Cấp 1 và 2, nên chất lượng sống cũng tăng, khả năng chi trả của khách hàng để sử dụng các sàn TMĐT cũng thoải mái hơn. - Hướng đến sự nhanh chóng, tiện lợi (thể hiện qua PreferredLoginDevice và NumberOfDeviceRegistered), không cần tham khảo quá nhiều do đã có mức độ quen thuộc nhất định với hệ thống.	- Khách hàng mới tham gia, có thể cân nhắc là nhóm khách hàng tiềm năng. - Tỉ lệ Nam khá cao so với Nữ nên ở cụm này, nhu cầu mua sắm sẽ tập trung nhiều về những sản phẩm về Công Nghệ. - Tuy vẫn phần đông là dân cư ở Cấp Thành Phố 1, nhưng đã có sự hiện diện nhiều hơn của Cấp Thành Phố 3, nên cụm 2 là những khách hàng có khả năng chi trả từ thấp tới vừa cho việc mua hàng trên sàn TMĐT. - Tâm lý của khách hàng ở cụm 2 hướng đến sự chắc chắn, kỹ càng hơn, khi họ cần phải tham khảo, xem xét, chọn lọc sản phẩm và thông tin sản phẩm nhiều hơn.

Bảng 22. Bảng so sánh đặc điểm riêng của 2 cụm

Biết được đặc điểm riêng của mỗi cụm như trên, với tư cách là một Nhà Quản Trị thì nhóm sẽ có những chính sách đề xuất khác nhau áp dụng riêng cho khách hàng ở từng cụm:

- Thực hiện thêm chính sách Hạng Thành Viên hay có những ưu đãi nhiều hơn cho khách hàng ở cụm 1 đồng thời các quảng cáo hay đề xuất mua hàng cũng đa dạng hơn (từ Công Nghệ đến Thời Trang). Còn ở cụm 2, sẽ có những hỗ trợ đặc biệt giúp cho người dùng dễ dàng tiếp cận với hệ thống (hướng dẫn từng bước, Pop-up Instructions,...), đồng thời phương thức quảng cáo cũng đơn giản, đánh vào trọng tâm nhiều hơn (tập trung vào một hay vài ngành hàng nhất định mà khách hàng ở cụm này quan tâm).
- Cần thực hiện đa dạng hóa trong cách tiếp cận khách hàng giữa các cụm, nếu như cụm 1 sẽ hướng đến sự chất lượng sản phẩm, sự tiện lợi, tiết kiệm thời gian tìm kiếm, thì ở cụm 2 sẽ hướng nhiều hơn đến sự đơn giản, thân thiện với khách hàng, đi cùng với chất lượng dịch vụ hỗ trợ khách hàng thường xuyên để từ đây vừa có thể tiếp tục duy trì quan hệ khách hàng với cụm 1 và có thêm nguồn khách hàng tiềm năng từ cụm 2.

3.5.2. Phát triển hệ thống quản lý quan hệ khách hàng (CRM)

Việc quản lý mối quan hệ với các khách hàng là một chiến lược kinh doanh nhằm gia tăng lợi nhuận bằng cách củng cố sự hài lòng, gia tăng lòng trung thành của khách hàng với thương hiệu. CRM sẽ có ích trong việc quản lý thông tin khách hàng hiệu quả

hơn, bao gồm thông tin cá nhân, lịch sử mua hàng, sở thích, yêu cầu và phản hồi từ khách hàng. Bên cạnh đó, CRM cho phép sàn thương mại điện tử theo dõi các thông tin về khách hàng và giao dịch của họ. Điều này giúp sàn thương mại điện tử có thể phát hiện và giải quyết các vấn đề của khách hàng nhanh chóng hơn và tạo cảm giác thoải mái, hài lòng cho khách hàng. Nhờ đánh giá những thông tin khách hàng, các sàn thương mại điện tử có thể tìm hiểu các xu hướng và phát triển các sản phẩm và dịch vụ mới tốt hơn để thu hút khách hàng. Ngoài ra thì, khi sử dụng CRM, sàn thương mại điện tử có thể cung cấp cho nhân viên dữ liệu khách hàng hoàn chỉnh và chi tiết, và đồng thời tương tác với khách hàng một cách chuyên nghiệp, tạo ra một mối quan hệ lâu dài và hợp tác tốt hơn, từ đó gia tăng sự tin tưởng của các nhóm khách hàng.

3.5.3. Phân tích tỷ lệ Customer Churn ngay khi nó xảy ra

Việc Phân tích tỷ lệ Customer Churn là cực kỳ quan trọng với sàn thương mại điện tử vì nó sẽ giúp đánh giá hiệu quả của chiến lược kinh doanh của doanh nghiệp cũng như tính toán số lượng khách hàng mới cần thu hút để bù đắp khách hàng đã rời bỏ dịch vụ. Nếu tỷ lệ rời bỏ quá cao, điều này có thể cho thấy những vấn đề về chất lượng sản phẩm/dịch vụ hoặc chiến lược marketing hiện tại chưa thật sự hiệu quả. Việc phân tích này giúp sàn thương mại điện tử nhanh chóng tìm ra các vấn đề và có kế hoạch xử lý kịp thời để giữ chân được khách hàng cũ và thu hút được khách hàng mới.

3.5.4. Cải thiện và nâng cao trải nghiệm mua sắm của khách hàng

Cải thiện và nâng cao trải nghiệm mua sắm của khách hàng trên sàn thương mại điện tử là một điều cần thiết vì mua sắm trực tuyến dần đã trở thành một phần của cuộc sống của rất nhiều người. Vì vậy, sàn thương mại điện tử cần phải đảm bảo rằng khách hàng có thể tìm thấy sản phẩm mình muốn và mua hàng một cách dễ dàng và nhanh chóng. Nếu khách hàng gặp vấn đề trong khi mua hàng trên sàn thương mại điện tử, họ có thể quay lại với các phương tiện mua sắm truyền thống và điều này có thể gây ra tổn thất cho doanh nghiệp. Bằng cách cải thiện trải nghiệm mua sắm của khách hàng, sàn thương mại điện tử có thể thu hút và giữ chân khách hàng hơn, tăng doanh số và tăng độ phát triển của doanh nghiệp.

CHƯƠNG 4: ĐÁNH GIÁ KẾT QUẢ CỦA MÔ HÌNH, KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

4.1. Tóm tắt

Từ dữ liệu Churn Rates, nhóm đã đưa ra được những yếu tố ảnh hưởng đến sự sụt giảm số lượng khách hàng sử dụng Hệ Thống Thương Mại Điện Tử và dự báo được những yếu tố này có ảnh hưởng thế nào đến biến phụ thuộc Churn. Thông qua việc lấy mẫu dữ liệu từ trang web https://www.kaggle.com/, thực hiện xử lý dữ liệu, huấn luyện

dữ liệu và phân lớp dữ liệu trên phần mềm Orange. Sau đó, chọn ra được mô hình Tree là mô hình phù hợp nhất và đưa ra được kết quả dự báo lưu dưới dạng E-Commerce Churn Dự báo.xlsx. Cũng từ phân lớp dữ liệu, nhóm đã tìm ra những vấn đề hiện đang có của hệ thống, sau đó đưa ra những giải pháp khả dĩ cho doanh nghiệp. Từ phân cụm dữ liệu, nhóm đã phân ra được 2 nhóm khách hàng có đặc điểm riêng biệt và đã kiến nghị những phương hướng phát triển trong tương lai.

4.2. Đánh giá

Nhóm đã hoàn thành được mục tiêu đề ra là dự báo những yếu tố ảnh hưởng đến sự sụt giảm khách hàng nhằm đưa ra những khuyến nghị phù hợp. Các lý thuyết ở chương 2 và chương 3 được nhóm vận dụng vào để xây dựng mô hình, phân tích độ khả năng rời bỏ của khách hàng, đưa ra dự báo có độ chính xác cao và chia cụm dữ liệu với số cụm hợp lý nhất.

4.3. Hướng phát triển

Tuy đã cố gắng hết sức để hoàn thành đồ án, nhóm vẫn không thể tránh khỏi những sai sót trong việc xử lý, phân tích và khai thác dữ liệu. Nhóm đã tự đánh giá và nhận ra những thiếu sót đó như là: quy mô dữ liệu khá lớn nên việc phân tích không thể hoàn toàn chính xác, tiền xử lý dữ liệu chưa tối ưu do có một lượng dữ liệu tương đối bị mất, nguồn của dữ liệu chưa rõ ràng,...

Vì thế, việc tìm hiểu sâu hơn về nguyên nhân, các yếu tố tác động hoặc liên quan đến vấn đề này là rất cần thiết. Các nghiên cứu tiếp theo có thể đi theo các hướng:

- Khám phá thêm các nhân tố khác có tác động đến ý định rời bỏ sử dụng Hệ Thống Thương Mại Điện Tử của khách hàng và đưa vào mô hình nghiên cứu để kiểm tra mức độ tác động của từng nhân tố. Các nhân tố khác có thể là chuẩn chủ quan, tệp khách hàng mục tiêu...
- Tiếp tục mở rộng quy mô, khảo sát trên số lượng khách hàng lớn hơn, và phải luôn cập nhật kết quả khảo sát theo từng năm, từng sàn thương mại điện tử cụ thể. Bởi vì xu hướng sử dụng của khách hàng luôn bị tác động và thay đổi dưới nhiều yếu tố khác nhau.